The ROC Isometrics Approach to Construct Reliable Classifiers

Stijn Vanderlooy[1], Ida G. Sprinkhuizen-Kuyper[2], Evgueni N. Smirnov[1], and H. Jaap van den Herik[1]

[1] MICC, Universiteit Maastricht, P.O. Box 616, 6200 MD Maastricht, The Netherlands, {s.vanderlooy,smirnov,herik}@micc.unimaas.nl
[2] NICI, Radboud University Nijmegen, P.O. Box 9104, 6500 HE Nijmegen, The Netherlands i.kuyper@nici.ru.nl

Abstract. We address the problem of applying machine-learning classifiers in domains where incorrect classifications have severe consequences. In these domains we propose to apply classifiers only when their performance can be defined by the domain expert prior to classification. The classifiers so obtained are called reliable classifiers. In the article we present three main contributions. First, we establish the effect on an ROC curve when ambiguous instances are left unclassified. Second, we propose the ROC isometrics approach to tune and transform a classifier in such a way that it becomes reliable. Third, we provide an empirical evaluation of the approach. From our analysis and experimental evaluation we may conclude that the ROC isometrics approach is an effective and efficient approach to construct reliable classifiers. In addition, a discussion about related work clearly shows the benefits of the approach when compared with existing approaches that also have the option to leave ambiguous instances unclassified.

Key words: ROC analysis, isometrics, abstaining classifiers, reliable classifiers, cost-sensitive classification

1 Introduction

In the past decades supervised learning algorithms have been applied to solve various classification tasks with growing success. Nonetheless, it remains difficult to apply the learned classifiers in domains such as law enforcement and medical diagnosis. These domains are characterized by high error costs indicating that incorrect classifications can have severe consequences. We propose to apply classifiers only if they are able to guarantee a preset classification performance on each class. The values of these performances are chosen in such a way that the costs of incorrect classifications that still may occur are acceptable. Since performance is defined prior to classification, the classifier becomes reliable and thus can be safely applied in the domain. The problem statement that we investigate therefore reads as follows: *Can we develop a feasible approach by which*

a classifier is constructed that guarantees a preset classification performance on each class? If the answer is affirmative, then we call the constructed classifier a *reliable classifier*.

Our answer to the problem statement focuses on binary classification tasks and it is decomposed into three stages. First, we analyze the effect on classification performance when instances with uncertainty in the true label are left unclassified. Second, we propose the ROC isometrics approach to fine-tune this effect in order to obtain a reliable classifier. Third, we use benchmark datasets to provide strong empirical evidence in order to verify that the ROC isometrics approach efficiently constructs reliable classifiers.

The motivation of the first stage is that the performance of a classifier is often deteriorated when instances that have uncertainty in the true label are classified. Therefore, we identify such uncertain instances and leave them unclassified. This procedure is called *abstention*. We use a Receiver Operating Characteristic (ROC) curve to determine the effect of abstention on the performance of a classifier. The result of abstention is a new ROC curve representing performance when abstention is applied. We analyze when and where this new ROC curve is better than the original one.

In the second stage a fine-tuning of abstention takes place. We wish to find those instances that should be left unclassified in order to obtain a desired reliable classifier. A trial-and-error procedure is clearly not feasible. Therefore, we propose an approach to find the minimum set of instances that should be left unclassified in such a way that the resulting classifier is reliable. Our approach to construct reliable classifiers, called the *ROC isometrics approach*, uses the classifier information provided by an ROC curve in combination with so-called ROC isometrics.

In the third stage, we provide an extensive empirical evaluation of the ROC isometrics approach. We test the approach by applying two popular classifiers on ten benchmark datasets. The datasets vary strongly in characteristics such as the size and the number of instances in each class. Experimental results confirm the correctness and efficiency of our approach. The advantages of the approach are shown and we discuss some issues that should be considered in practice.

The remainder of the article is as follows. In Section 2 we provide some background knowledge. An analysis of abstention (the first stage) is given in Section 3. In Section 4 we explain how isometrics are used in our approach to construct reliable classifiers (the second stage) and we provide a formal analysis to show the validity of the approach. The experimental verification (the third stage) is given in Section 5. Related work is discussed in Section 6. In Section 7 we present our conclusions. The proofs of all theorems are excluded from the main text and gathered in the appendix. This article uses and extends the work in [1, 2].

2 Background

In this section we review the concepts used throughout the article. In Subsection 2.1 we consider two types of classifiers and some basic properties of classification performance. In Subsection 2.2 we show the concept of an ROC curve as a natural visualization tool for thresholding. In Subsection 2.3 we review a general scheme for performance evaluation. In Subsection 2.4 we discuss classifiers that can abstain from classifying instances.

2.1 Classifiers and Thresholding

Many real-life problems can be formulated as binary classification tasks in which the classes are indicated as positive (p) and negative (n). We mention disease diagnosis, fraud detection, and relevant document retrieval. The supervised machine-learning setting is considered to solve these tasks. Below we discuss two types of classifiers and the notion of thresholding scores.

Discrete Classifier A *discrete classifier* is a mapping from instances to classes. The outcome of a classification can be a true positive, false positive, true negative, and false negative. The total number of these outcomes are denoted by TP, FP, TN, and FN, respectively. Henceforth, they are called *performance statistics*. The number of positive instances is $P = TP + FN$ and the number of negative instances is $N = TN + FP$. From these numbers we derive:

$$tpr = \frac{TP}{TP + FN} \text{ , and } tnr = \frac{TN}{TN + FP} \text{ ,} \tag{1}$$

$$fpr = \frac{FP}{FP + TN} \text{ , and } fnr = \frac{FN}{TP + FN} \text{ ,} \tag{2}$$

where true positive rate is denoted by tpr and true negative rate by tnr.[3] False positive rate and false negative rate are denoted by fpr and fnr, respectively.

Scoring Classifier In contrast to a discrete classifier, a *scoring classifier* outputs two non-negative values $\hat{\mathbb{P}}(x \mid p)$ and $\hat{\mathbb{P}}(x \mid n)$ that represent the degree that an instance x is positive and negative, respectively. These values can be estimated probabilities or simply uncalibrated values with the only property that higher values indicate a higher probability. The *score* of an instance is then defined as:

$$l(x) = \frac{\hat{\mathbb{P}}(x \mid p)}{\hat{\mathbb{P}}(x \mid n)} \text{ ,} \tag{3}$$

or infinity in case of $\hat{\mathbb{P}}(x \mid n) = 0$. Scores can be used to rank instances from most likely positive to most likely negative [3]. Applying a numerical threshold

[3] Depending on the research field tpr is sometimes called positive accuracy, recall, or sensitivity. Negative accuracy and specificity are synonyms for tnr.

on scores transforms a scoring classifier into a discrete classifier. An instance is classified as positive if its score is higher than or equal to the chosen threshold, and otherwise as negative.

Thresholding Scores A scoring classifier is almost never optimal, i.e., there will exist negative instances that received a higher score than some positive instances. In other words, the produced scores do not allow to distinguish correctly between the positive class and the negative class. Therefore, applying a threshold often results in (unaffordable) incorrect classifications.

Clearly, the threshold determines the trade-off between the number of false positives and false negatives. Decreasing the threshold decreases FN and increases TP. Simultaneously, TN decreases and FP increases. Increasing the threshold results in the opposite movement of the performance statistics. In the following subsection we use the ROC curve to visualize this trade-off.

2.2 ROC Curves

The *Receiver Operating Characteristic curves* (ROC curves) have recently been introduced in the field of machine learning for visualizing, selecting, and combining classifiers [4–7]. In this article we focus on visualization and selection.

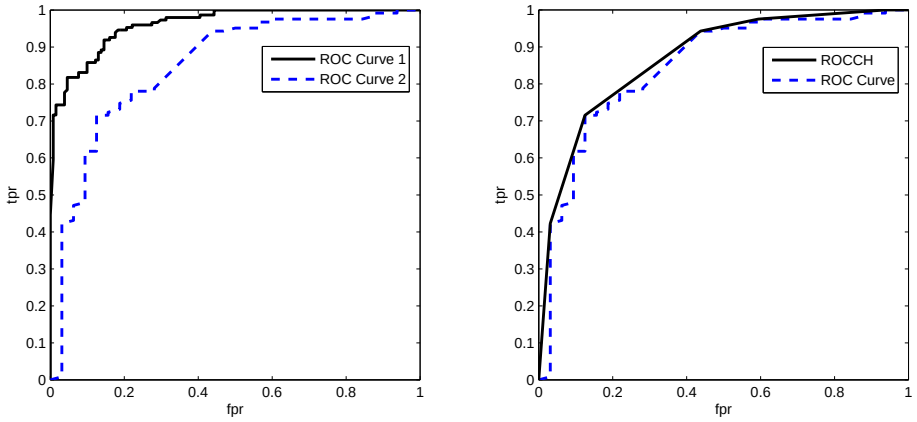
An ROC curve is defined for a scoring classifier. It is a plot with two dimensions: fpr on the horizontal axis and tpr on the vertical axis. The (fpr, tpr) plane is called ROC space. Implicitly, the ROC space also includes tnr and fnr since $tnr = 1 - fpr$ and $fnr = 1 - tpr$. Given a scoring classifier, each threshold on the score results in a discrete classifier that corresponds to exactly one point in ROC space. The connection of points obtained by applying all possible thresholds in decreasing order is the ROC curve. Note that the ROC space is independent of the cost distribution and the class distribution governing the application domain (more on this in the next subsection).

Figure 1(a) shows two ROC curves of which one is said to be dominating. For each value of fpr, the dominating ROC curve has an equal or higher value of tpr than the ROC curve being dominated [8]. Thus, for fixed fpr, the discrete classifier constructed from the dominating ROC curve is at least as good as the one constructed from the other curve. In most practical applications, there is no ROC curve that dominates. Instead, the curves intersect each other and a curve is said to dominate in one or more regions of ROC space [9].

An empirical ROC curve often has concavities. Its convex hull, denoted by ROCCH, removes the concavities as follows. Instances in a concavity can be ranked arbitrarily by assigning them equal scores. The expected performance is shown by averaging the optimistic and pessimistic ranking of these instances [5]. Figure 1(b) shows an ROC curve and corresponding ROCCH.

Theorem 1. *For any point on or below the ROCCH a classifier can be constructed by thresholding the scores in such a way that it achieves the performance statistics represented by that point.*

The proof of this theorem for the case of a point on the ROCCH is given in [6]. For the interested reader we included our own proof in the appendix. The case of a point below the convex hull is a straightforward extension of interpolating classifiers on the convex hull [10]. For simplicity of presentation we will assume that ROC curves are convex and all points can be obtained by a threshold.



(a) (b)

Fig. 1. An ROC curve shows the performance of a scoring classifier: (a) ROC curve 1 dominates ROC curve 2, and (b) an ROC curve and corresponding ROCCH. Note that an empirical curve is always dominated by its convex hull.

2.3 Skew-Sensitive Performance Evaluation

Most of the widely-used performance metrics do not consider the *cost distribution* and the *class distribution* of the application domain, hereby neglecting the central role of these two distributions in performance evaluation [9, 11–14]. For instance, accuracy assumes that the class distribution is constant and relatively balanced. Optimizing classifiers in terms of accuracy when the class distribution is very skewed often prefers trivial classifiers that ignore the minority class. Large class skews are however common and the minority class is often the class of interest. Also, accuracy assumes that the costs of a false negative and a false positive are equal. This assumption is unrealistic since in practice a false positive is often more (or less) expensive than a false negative. For example, consider law enforcement applications where a false positive can be a person who is sentenced while not guilty, whereas a false negative is a guilty person who is not sentenced.

Clearly, the cost distribution and the class distribution should both be taken into account when evaluating classifiers. For this reason, the *skew ratio* is defined

to express the relative importance of the negative class versus the positive class [15, 16]:

$$c = \frac{c(p,n)}{c(n,p)} \frac{\pi_n}{\pi_p} \ , \tag{4}$$

where $c(p,n)$ and $c(n,p)$ are the costs of a false positive and a false negative, respectively. The probabilities of a negative instance and positive instance are denoted by $\pi_n = \frac{N}{P+N}$ and $\pi_p = \frac{P}{P+N}$. The class distribution is $\frac{\pi_n}{\pi_p} = \frac{N}{P}$. The cost distribution is given by the costs of incorrect classifications only since benefits of true positives and true negatives can be incorporated by substracting them from the errors. This operation normalizes the cost matrix in such a way that the two values on the main diagonal are zero [15, 17]. In the remainder of the article we therefore always assume zero costs for correct classifications.

The skew ratio can be plugged as a parameter into a performance metric. This allows us to measure classification performance in any possible scenario of cost distributions and class distributions. For instance, we may set the skew ratio equal to the class distribution in the test set if we believe that this is representative for the application domain. If we believe that this is not the case, then we may insert the correct class distribution into the skew ratio. The same holds for the cost distribution. In general, if $c < 1$ then the positive class is most important and if $c > 1$ then the negative class is most important. In the following we set c equal to the class distribution of the test set (i.e., $c = \frac{N}{P}$) to keep equations and derivations easy to comprehend. The reader should keep in mind that this is without any restriction: our results are also valid when a different class distribution or cost distribution is incorporated.

2.4 Abstaining Classifiers

In complex real-life applications, a classifier often encounters instances to be classified that are different from instances encountered during the training phase. In particular when incorrect classifications have high costs, it is desired to abstain from classifying instances for which there is uncertainty in the true label. Henceforth, we refer to these instances as *uncertain instances*.

An *abstaining classifier* is a classifier that can abstain from classifying uncertain instances [18–24, 17]. Such a classifier can improve the classification performance significantly, even for low abstention rates. It simulates the behaviour of human experts. For example, in medical diagnosis an expert does not state a possibly incorrect diagnosis but she says "I do not know" and performs more tests. Uncertain instances that are left unclassified can be: (1) rejected from the system, (2) passed to a human for classification, or (3) classified by another (possibly more expensive) classifier [25–27].

Uncertain instances are instances with scores that do not clearly indicate the correct classification. Therefore, an abstaining classifier can be implemented as a *reject rule* based on scores. We consider the case where the reject rule uses two thresholds to identify uncertain instances. The classification scheme is presented in Algorithm 1 where thresholds $a > b$ are used to classify evidently

Algorithm 1: Classification scheme of abstaining classifiers

input : Instance x to be classified, thresholds $a > b$
output: Classification for x

$l(x) \leftarrow \frac{\hat{\mathbb{P}}(x|p)}{\hat{\mathbb{P}}(x|n)}$

if $l(x) \geq a$ **then**
  return p
else if $l(x) \leq b$ **then**
  return n
else
  return u

positive instances and evidently negative instances, respectively. The remaining instances are left unclassified (represented by class u). Note that the condition $a = b$ would imply an abstaining classifier with no abstention rate, i.e., the condition implies a common discrete classifier.

We denote the number of positive instances left unclassified by UP and the number of negative instances left unclassified by UN. Decreasing the number of unclassified instances is achieved by: (1) decreasing the value of a, (2) increasing the value of b, or (3) both. Decreasing the value of a results in fewer unclassified positive instances but also in more false positives. Increasing the value of b results in fewer unclassified negative instances but also in more false negatives. What to prefer depends on the requirements of the application domain. The proportion of abstention is defined by the *unclassified positive rate* (upr) and the *unclassified negative rate* (unr):

$$upr = \frac{UP}{TP + FN + UP} \text{ , and} \tag{5}$$

$$unr = \frac{UN}{FP + TN + UN} \text{ .} \tag{6}$$

The goal of abstaining classifiers is to keep the total number of unclassified instances as low as possible while still guaranteeing a significantly higher performance than that obtained by a classifier without the reject rule implemented.

3 Effect of Abstention in ROC Space

In this section we provide our results on how abstention can be visualized and analyzed in ROC space (the first stage). In Subsection 3.1 we describe the effect on the ROC curve of a classifier when it is transformed into an abstaining classifier. The so-called dominance relations between the original ROC curve and that of the abstaining classifier are provided in Subsection 3.2.

3.1 Abstention ROC Curves

Consider a scoring classifier and its ROC curve. Henceforth, we call this curve the *original curve*. Each point on this curve is the result of applying some threshold on the scores of instances. To transform the scoring classifier into an abstaining classifier we have to define two thresholds a and b. These thresholds can be used to construct the ROC curve of the abstaining classifier by using the classified instances only. We call this curve the *abstention curve*.

We write (fpr_a, tpr_a) and (fpr_b, tpr_b) to denote the points on the original curve that correspond to the thresholds a and b of the abstaining classifier (see Fig. 2 for an illustration). Intuitively, the abstention curve is obtained by only considering thresholds for which the resulting false positive rate is at most fpr_a or at least fpr_b. This corresponds with not covering the part of the original curve between the points (fpr_a, tpr_a) and (fpr_b, tpr_b) since the construction of this part involved instances that are now left unclassified. Hence, we define the *uncovered part* of the abstaining classifier as the part between (fpr_a, tpr_a) and (fpr_b, tpr_b). The *covered part* is defined as the part from $(0, 0)$ to (fpr_a, tpr_a) and from (fpr_b, tpr_b) to $(1, 1)$. It follows that the covered part is the complement of the uncovered part. By definition, the unclassified positive rate and the unclassified negative rate are described as follows:

$$upr = tpr_b - tpr_a \text{ , and} \tag{7}$$

$$unr = fpr_b - fpr_a \text{ .} \tag{8}$$

The transformation from the original curve to the abstention curve is given in Theorem 2. The key idea is that a point on the covered part of the original curve is associated with a point on the abstention curve such that the corresponding discrete classifiers classify the same positive and negative instances. More details can be found in the proof (see appendix).

Theorem 2. *If the part between (fpr_a, tpr_a) and (fpr_b, tpr_b) of an ROC curve is not covered, and $0 < upr < 1$ and $0 < unr < 1$, then points (fpr_i, tpr_i) on this curve between $(0, 0)$ and (fpr_a, tpr_a) are transformed into points (fpr'_i, tpr'_i) on the abstention ROC curve such that:*

$$\begin{aligned} fpr'_i &= \frac{fpr_i}{1 - unr} \text{ , and} \\ tpr'_i &= \frac{tpr_i}{1 - upr} \text{ .} \end{aligned} \tag{9}$$

Also, points (fpr_i, tpr_i) between (fpr_b, tpr_b) and $(1, 1)$ are transformed into points (fpr'_i, tpr'_i) on the abstention ROC curve such that:

$$\begin{aligned} fpr'_i &= 1 - \frac{1 - fpr_i}{1 - unr} \text{ , and} \\ tpr'_i &= 1 - \frac{1 - tpr_i}{1 - upr} \text{ .} \end{aligned} \tag{10}$$

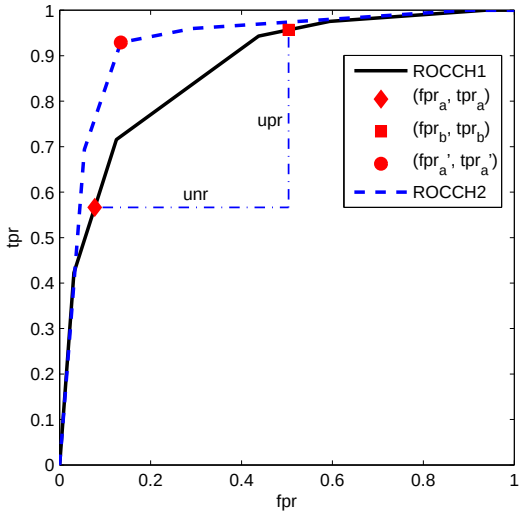


Fig. 2. Abstaining classifier in ROC space: the abstention ROCCH (– – –) is obtained by not covering the part between points (fpr_a, tpr_a) and (fpr_b, tpr_b) of the original ROCCH (——). The length of the horizontal and vertical dash-dotted lines equals *unr* and *upr*, respectively.

Note that points (fpr_a, tpr_a) and (fpr_b, tpr_b) are transformed into the same point on the abstention ROC curve. Figure 2 shows an example of a transformation with $upr = 0.34$ and $unr = 0.41$. The transformation of the endpoints of the uncovered part is also indicated. It is clear from the figure that the effect of abstention on classification performance is specific to the dataset and the classifier used. The ROCCHs presented in this article will reflect a classifier with moderate performance allowing us to emphasize the effect of abstention on the curves. In general, the better the scores of a classifier, the less need for a reject rule. In other words, the effect of abstention becomes smaller with ROCCHs that have a few line segments close to the optimal point $(0, 1)$.

Theorem 3. *If the original ROC curve is convex, then the ROC curve of the abstaining classifier is also convex.*

Theorem 3 shows that we can restrict ourselves to convex ROC curves in the remainder of the article. In the next subsection we will introduce the so-called dominance relations between the (convex) original curve and corresponding (convex) abstention curve.

3.2 Analysis of Dominance Relations

In this subsection we will compare the original ROCCH with the abstention ROCCH by means of *dominance* relations. If the abstention ROCCH dominates

in a region of ROC space, then abstention yields the highest performance when applied in the corresponding scenario of cost distributions and class distributions [5]. Clearly, we wish that the abstention ROCCH dominates as much as possible since this implies that abstention: (1) gives a better performance than was previously possible, and (2) is robust to changes in the application domain.

Theorem 4 shows that the abstention convex hull always dominates the original one on the part from $(0,0)$ to (fpr_a, tpr_a) if the unclassified negative rate is less than or equal to the unclassified positive rate. Theorem 5 provides a condition for dominance on the part from (fpr_b, tpr_b) to $(1,1)$. Both theorems conform with our intuition.

Theorem 4. *If $unr \leq upr$ then the abstention ROCCH dominates the original one on the part from $(0,0)$ to (fpr_a, tpr_a).*

Theorem 5. *If $unr \geq upr$ then the abstention ROCCH dominates the original one on the part from (fpr_b, tpr_b) to $(1,1)$.*

Illustrations of Theorem 4 and Theorem 5 are given in Figs. 3(a) and 3(b), respectively. Since it is always true that $unr \leq upr$ or $unr \geq upr$, at least one of the two theorems is applicable. Thus, there is always a part where the abstention ROCCH is dominating. Moreover, we have the following corollary.

Corollary 1. *If $unr = upr$ then the abstention ROCCH dominates the original one on the covered part.*

This corollary is important in practice, in particular when the cost distribution and the class distribution: (1) cannot be estimated precisely, and/or (2) are subject to strong change.[4] Corollary 1 guarantees that the abstaining classifier outperforms the original classifier for all possible cost and class distributions.

Theorems 6 and 7 give conditions for dominance on the covered part when the first line segment of the original ROCCH is vertical and the last line segment is horizontal, respectively. Both theorems will prove useful to define new conditions for dominance on the covered part.

Theorem 6. *If the original ROCCH contains a point $(0, tpr_0)$ with $tpr_0 > 0$, then the abstention ROCCH dominates the original one on the covered part if $unr > upr$ and $tpr_a \leq \frac{1}{1-\frac{upr}{unr}} tpr_0$.*

Theorem 7. *If the original ROCCH contains a point $(tnr_0, 1)$ with $tnr_0 = 1 - fpr_0 > 0$, then the abstention ROCCH dominates the original one on the covered part if $upr > unr$ and $1 - fpr_b \leq \frac{1}{1-\frac{unr}{upr}}(1 - fpr_0)$.*

Figure 4(a) shows a situation where Theorem 7 applies. Combining Theorems 6, 7, and Corollary 1 results in a relaxation of Corollary 1. A relaxation is desired since the condition in the corollary can be difficult to satisfy in practice. Our relaxation is given in Corollary 2 and is illustrated in Fig. 4(b).

[4] At first sight it may seem that real-life domains possessing continuously evolving class and cost distributions are exceptional. However, these domains actually occur often. Some representative examples can be found in [11].

Corollary 2. *If the original ROCCH contains two points* $(0, tpr_0)$ *and* $(tnr_0, 1)$ *with* $tpr_0 > 0$ *and* $tnr_0 = 1 - fpr_0 > 0$*, and if* $\frac{fpr_0 - fpr_b}{1 - fpr_b} \leq \frac{unr}{upr} \leq \frac{tpr_a}{tpr_a - tpr_0}$*, then the abstention ROCCH dominates the original one on the covered part.*

This corollary states that the abstention ROCCH dominates on the covered part when the first line segment is vertical, the last line segment is horizontal, and the unclassified positive rate is approximately the unclassified negative rate. Corollary 2 is easier to satisfy in practice than Corollary 1. Indeed, most scoring classifiers do not assign the highest score to a negative instance and the lowest score to a positive instance. Therefore, the corresponding ROCCHs have a first line segment that is vertical and a last line segment that is horizontal. The lengths of these two line segments may be small, but nonetheless Corollary 2 applies when we define an abstaining classifier such that $upr \approx unr$.

4 How to Construct Reliable Classifiers

In this section we introduce our approach to construct reliable classifiers, i.e., classifiers that guarantee at least a classification performance as preset by the domain expert (the second stage). For this purpose we first discuss ROC isometrics in Subsection 4.1. Afterwards we present and analyze our approach in Subsections 4.2 and 4.3, respectively. From the analysis we conclude in Subsection 4.4 that the approach is indeed a solution to our problem statement.

4.1 ROC Isometrics

The main tools of our approach to construct reliable classifiers are *ROC isometrics*, which are defined as curves in ROC space that connect points with the same value for a (skew-sensitive) performance metric [16, 28]. We consider the following three performance metrics: (1) precision, (2) F-measure, and (3) m-estimate. Each metric has a positive and a negative variant. The positive variant measures performance on the positive classifications and the negative variant measures performance on the negative classifications. Hence, in total we consider six different performance metrics. Table 1 summarizes these metrics and Table 2 shows the corresponding isometrics. An isometric is obtained by rewriting the equation of the metric to that of a curve in ROC space. Note that all isometrics considered in this article are linear curves. For generality, the isometric of a positive (negative) variant of a performance metric is called a positive (negative) isometric.

Exactly one isometric corresponds to a specific skew ratio and a specific value for the underlying performance metric. The isometric contains all points which have the specified performance value under the conditions defined by the skew ratio. Varying the skew ratio or performance value results in isometrics that rotate around a single point in which the performance metric is undefined. For each performance metric we will now investigate its isometrics in more detail.

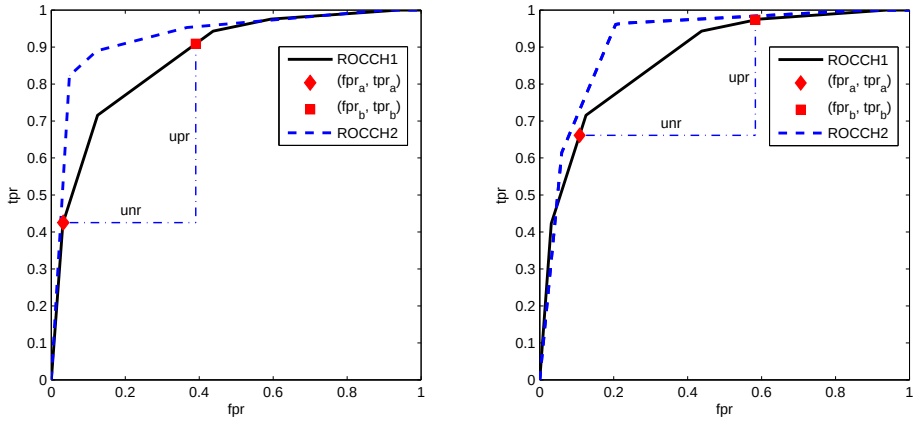


Fig. 3. Dominating abstention ROCCH (– – –) in a specific ROC region: (a) application of Theorem 4, and (b) application of Theorem 5.

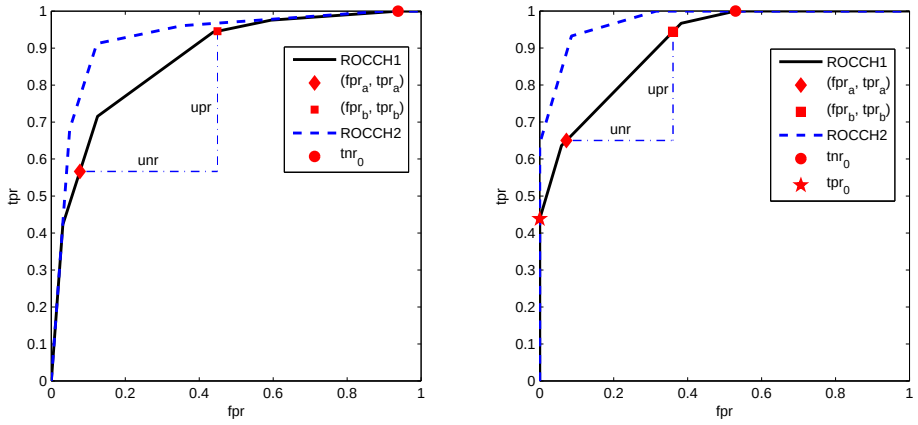


Fig. 4. Dominating abstention ROCCH (– – –) on the covered part: (a) application of Theorem 7, and (b) application of Corollary 2.

Table 1. Skew-sensitive performance metrics defined in terms of fpr, tpr, $c = \frac{N}{P}$, $\alpha \in \mathbb{R}^+$, and $\hat{m} = \frac{m}{P+N}$. For each of the conventional metrics we introduce a positive variant and a negative variant.

Metric	Indicator	Formula
Positive precision	$prec_p^c$	$\frac{tpr}{tpr+c\ fpr}$
Negative precision	$prec_n^c$	$\frac{tnr}{tnr+\frac{1}{c}fnr}$
Positive F-measure	$F_p^{c,\alpha}$	$\frac{\left(1+\alpha^2\right)tpr}{\alpha^2+tpr+c\ fpr}$
Negative F-measure	$F_n^{c,\alpha}$	$\frac{\left(1+\alpha^2\right)tnr}{\alpha^2+tnr+\frac{1}{c}fnr}$
Positive m-estimate	$mest_p^{c,\hat{m}}$	$\frac{tpr+\hat{m}}{tpr+c\ fpr+\hat{m}(1+c)}$
Negative m-estimate	$mest_n^{c,\hat{m}}$	$\frac{tnr+\hat{m}}{tnr+\frac{1}{c}fnr+\hat{m}\frac{1+c}{c}}$

Table 2. ROC Isometrics defined in terms of fpr, tpr, $c = \frac{N}{P}$, $\alpha \in \mathbb{R}^+$, and $\hat{m} = \frac{m}{P+N}$. Each of the six ROC isometrics is defined by a skew ratio and a value for the underlying performance metric.

Metric	Isometric
Positive precision	$tpr = \frac{prec_p^c}{1-prec_p^c}\ c\ fpr$
Negative precision	$tpr = \frac{1-prec_n^c}{prec_n^c}\ c\ fpr + 1 - \frac{1-prec_n^c}{prec_n^c}\ c$
Positive F-measure	$tpr = \frac{F_p^{c,\alpha}}{1+\alpha^2-F_p^{c,\alpha}} c\ fpr + \frac{\alpha^2\ F_p^{c,\alpha}}{1+\alpha^2-F_p^{c,\alpha}}$
Negative F-measure	$tpr = \frac{1+\alpha^2-F_n^{c,\alpha}}{F_n^{c,\alpha}} c\ fpr + 1 + \frac{(1+\alpha^2)\left(F_n^{c,\alpha}-1\right)}{F_n^{c,\alpha}} c$
Positive m-estimate	$tpr = \frac{mest_p^{c,\hat{m}}}{1-mest_p^{c,\hat{m}}} c\ fpr + \frac{\hat{m}\left(mest_p^{c,\hat{m}}(1+c)-1\right)}{1-mest_p^{c,\hat{m}}}$
Negative m-estimate	$tpr = \frac{1-mest_n^{c,\hat{m}}}{mest_n^{c,\hat{m}}} c\ fpr + 1 - \frac{1-mest_n^{c,\hat{m}}}{mest_n^{c,\hat{m}}} c + \frac{\hat{m}\left(mest_n^{c,\hat{m}}(1+c)-c\right)}{mest_n^{c,\hat{m}}}$

Precision Positive precision is defined as the proportion of true positives to the total number of positive classifications:

$$prec_p^c = \frac{TP}{TP+FP} = \frac{tpr}{tpr + c\ fpr}\ . \tag{11}$$

The isometrics are lines that rotate around point $(0,0)$. Intuitively, the higher the skew ratio and/or the positive precision value, the more the isometric approaches the optimal point $(0,1)$.

Negative precision is defined as the proportion of true negatives to the total number of negative classifications:

$$prec_p^c = \frac{TN}{TN+FN} = \frac{tnr}{tnr + \frac{1}{c}\ fnr}\ . \tag{12}$$

Corresponding isometrics rotate around point $(1,1)$. In addition, lower skew ratios and/or higher negative precision values result in isometrics approaching the optimal point $(0,1)$.

Figure 5 shows $prec_p^c$-isometrics and $prec_n^c$-isometrics for $c = 1$. In this figure and in the subsequent figures, the value of the performance metric is varied from 0.1 to 0.9 in steps of 0.1. The described movement of the precision isometrics with regard to the skew ratio and the performance value holds for all other isometrics as well.

***F*-measure** Positive precision is maximized when all positive classifications are correct. Hence, maximal positive precision can be obtained by correctly classifying at least one evidently positive instance while making no false positives. It is recommended to combine the metric with the true positive rate in order to know whether it uses sufficient positive instances to be considered as accurate. For this reason the positive F-measure was introduced [29]:

$$F_p^{c,\alpha} = \frac{(1+\alpha^2)\ prec_p^c\ tpr}{\alpha^2\ prec_p^c + tpr} = \frac{\left(1+\alpha^2\right) tpr}{\alpha^2 + tpr + c\ fpr} \quad , \tag{13}$$

where parameter α indicates the importance given to $prec_p^c$ relative to tpr. In general, if $\alpha < 1$ ($\alpha > 1$) then tpr is less (more) important than $prec_p^c$. If $\alpha = 1$ then they are equally important. The positive F-measure has the nice property that it is high when both $prec_p^c$ and tpr are high.[5] Its isometrics are lines that rotate around point $(-\alpha^2/c, 0)$ and therefore they can be seen as a shifted version of the $prec_p^c$-isometrics. The larger c and/or the smaller α, the smaller the difference between $F_p^{c,\alpha}$-isometrics and $prec_p^c$-isometrics.

Negative precision has a similar disadvantage as positive precision, i.e., it can be trivially maximized. Analogously, the negative F-measure is used for the trade-off between $prec_n^c$ and tnr:

$$F_n^{c,\alpha} = \frac{(1+\alpha^2)\ prec_n^c\ tnr}{\alpha^2\ prec_n^c + tnr} = \frac{\left(1+\alpha^2\right) tnr}{\alpha^2 + tnr + \frac{1}{c}\ fnr} \quad , \tag{14}$$

where parameter α now indicates the importance given to $prec_n^c$ relative to tnr. Corresponding isometrics are shifted versions of the $prec_n^c$-isometrics and rotate around point $(1, 1+\alpha^2 c)$. The smaller c and/or the smaller α, the less difference we observe with $prec_n^c$-isometrics.

Figure 6 shows $F_p^{c,\alpha}$-isometrics and $F_n^{c,\alpha}$-isometrics for $c = 1$ and $\alpha = 1$ in the relevant region $(0,1) \times (0,1)$ of ROC space.

***m*-estimate** The m-estimate computes a precision assuming that m imagined instances are classified a priori. The two main reasons why it is favoured over the conventional precision are: (1) less sensitive to noise, and (2) more effective

[5] Note that $prec_p^c$ and tpr are typically antagonistic: if $prec_p^c$ goes up, then tpr usually goes down, and vice versa.

in avoiding overfitting [30] [31, Ch. 8-10]. This is particularly true when the class distribution is highly skewed and the classification performance for the minority class is measured. By including a prior we correct for an overly optimistic performance when too few data are available [28].

The positive m-estimate assumes that m instances are a priori classified as positive. These instances are distributed according to the prior class distribution. We therefore have the following definition:

$$mest_p^{c,m} = \frac{TP + m\frac{P}{P+N}}{TP + FP + m} = \frac{tpr + \frac{m}{P+N}}{tpr + c\,fpr + \frac{m}{P}} \quad . \tag{15}$$

To eliminate the absolute numbers P and N, we define $\hat{m} = \frac{m}{P+N}$ and obtain the formula given in Table 1. Corresponding isometrics rotate around $(-\hat{m}, -\hat{m})$. If $\hat{m} = 0$ then $prec_p^c$-isometrics are obtained. If $\hat{m} \to \infty$ then the performance metric converges to $\frac{1}{1+c} = \mathbb{P}(p)$ and the corresponding isometric is the ascending diagonal.

The case of the negative m-estimate is similar. A small calculation shows that the rotation point of the isometrics is $(1 + \hat{m}, 1 + \hat{m})$. Figure 7 shows $mest_p^{c,\hat{m}}$-isometrics and $mest_n^{c,\hat{m}}$-isometrics for $c = 1$ and $\hat{m} = 0.1$.

4.2 ROC Isometrics Approach

Our approach departs from a scoring classifier and its ROCCH. The inputs from the domain expert are: (1) the skew ratio, (2) the desired performance on the positive classifications, and (3) the desired performance on the negative classifications. Hence, a positive isometric and a negative isometric can be constructed. The intersection point of the positive isometric and the ROCCH represents by definition a classifier with the desired performance on the positive class. We recall that we denote this point by (fpr_a, tpr_a). Analogously, the intersection point (fpr_b, tpr_b) of the negative isometric and the ROCCH represents a classifier with the desired performance on the negative class.[6] The intersection point of the isometrics themselves represents the reliable classifier. Dependent on the location of this intersection point we distinguish the following three cases as shown in Fig. 8.

- **Case 1:** the isometrics intersect on the ROCCH.
  By definition we have that the reliable classifier lies on the ROCCH. It follows from Theorem 1 that the classifier can be constructed by a single threshold that is applied on the score of instances.

[6] Here, we assume that the empirical ROCCH is an accurate estimate of the true ROCCH. This assumption can be verified by deriving confidence bands along the curve [32]. In essence, the lower band and the upper band can also be used as a guideline to adjust the thresholds when the assumption does not seem to hold. However, from our experiments, we found that even for small datasets the empirical curve is accurate enough. Recent generalization bounds for the area under the ROC curve give further evidence for this observation [33].

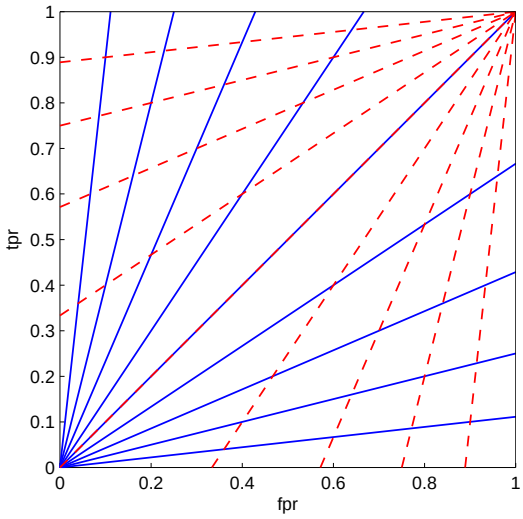


Fig. 5. Precision isometrics: $prec_p^1$-isometrics (——) and $prec_n^1$-isometrics (– – –).

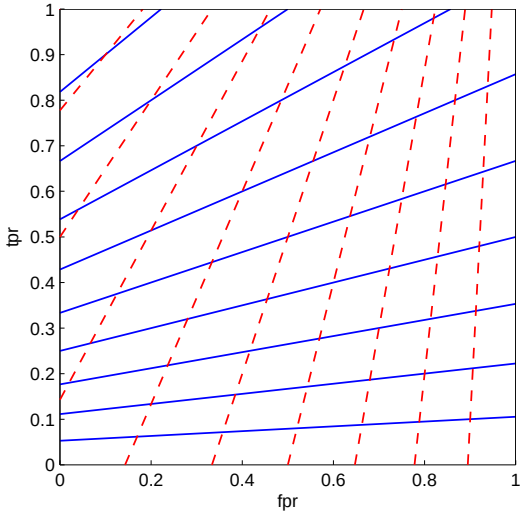


Fig. 6. F-measure isometrics: $F_p^{1,1}$-isometrics (——) and $F_n^{1,1}$-isometrics (– – –).

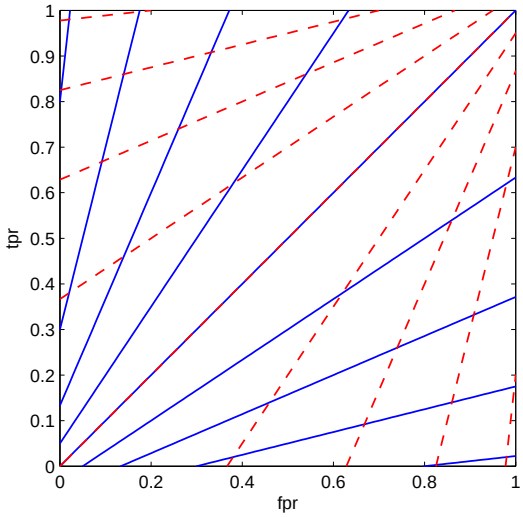
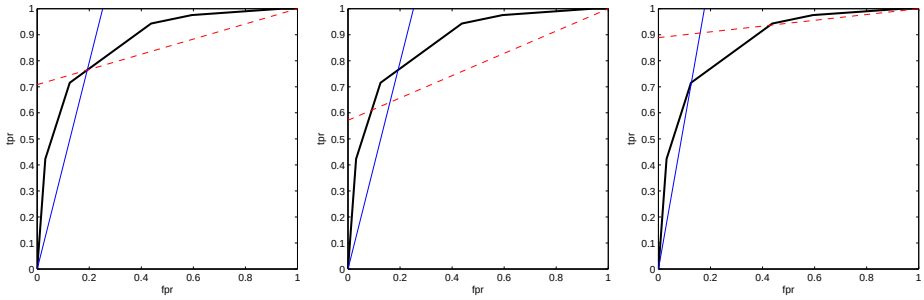


Fig. 7. m-estimate isometrics: $mest_p^{1,0.1}$-isometrics (——) and $mest_n^{1,0.1}$-isometrics (– – –).

- **Case 2:** the isometrics intersect below the ROCCH.
  Theorem 1 also applies in this case. However, classifiers corresponding to points on the ROCCH between (fpr_b, tpr_b) and (fpr_a, tpr_a) have a higher performance on both classes. Thus, one of these classifiers should be preferred since we are not interested in downgrading a performance.
- **Case 3:** the isometrics intersect above the ROCCH.
  The reliable classifier cannot be constructed using a single threshold. Our proposed solution is to define an abstaining classifier to filter out instances for which there is uncertainty in the true label. The thresholds $a > b$ of the abstaining classifier are identified to correspond with points (fpr_a, tpr_a) and (fpr_b, tpr_b), respectively. When there is more than one intersection point for the positive (negative) isometric and the ROCCH, the intersection point with highest tpr (lowest fpr) is chosen such that $fpr_a < fpr_b$. In this way the number of unclassified instances is minimized. Theorem 2 can then be used to construct the abstention ROCCH. Consequently, the intersection points are transformed to the same point $(fpr'_a, tpr'_a) = (fpr'_b, tpr'_b)$ on the abstention ROCCH. In the next subsection we show that the type of isometrics (precision, F-measure, and m-estimate) defines the effect on the performance of the classifier corresponding to (fpr'_a, tpr'_a).

The three cases that we distinguished cover all possible scenarios. Cases 1 and 2 show that the reliable classifier can easily be constructed since the desired performances as defined by the domain expert can be obtained by a non-abstaining classifier. An interpretation of case 3 is however not straightforward. In the next



(a) (b) (c)

Fig. 8. Location of the intersection point between a positive isometric and a negative isometric: (a) Case 1, (b) Case 2, and (c) Case 3.

subsection we present an analysis to verify whether the abstaining classifier as proposed in case 3 is the reliable classifier.

4.3 Analysis of the ROC Isometrics Approach

We provide a separate analysis for each performance metric since different performance metrics imply different isometrics. For brevity and clarity, the proofs of the formal analysis (gathered in the appendix) consider the case where the skew ratio is defined as the class distribution. The results generalize in a straightforward way to the case in which the skew ratio also incorporates the cost distribution, i.e., $c = \frac{c(p,n)}{c(n,p)} \frac{N}{P}$.

Precision Theorem 8 shows the existence of a reliable classifier when the desired performance is represented by precision on each class. Theorem 9 shows that the ROC isometrics approach can also construct a reliable classifier with a preset accuracy. This is achieved by constructing two precision isometrics with the desired accuracy as the performance value.

Theorem 8. *If the points* (fpr_a, tpr_a) *and* (fpr_b, tpr_b) *are defined by a* $prec_p^c$*-isometric and* $prec_n^c$*-isometric respectively, then the point* (fpr'_a, tpr'_a) *has the precisions of both isometrics.*

Theorem 9. *If point* (fpr'_a, tpr'_a) *has* $prec_p^c = prec_n^c$*, then the accuracy in this point equals the precisions.*

Note that Theorem 9 implies that: (1) the ROC isometrics approach can guarantee a preset accuracy on each class, and (2) the cost distribution and the class distribution can be incorporated via the skew ratio. Therefore, the approach overcomes the two problems of the conventional performance evaluation by means of accuracy, as explained in Subsection 2.3. From the proof of Theorem 9 it follows that, if the precisions of the isometrics are not equal, then the accuracy is bounded by the smallest precision and largest precision.

***F*-measure** Theorem 10 shows the existence of a reliable classifier when the preset performance is represented by F-measure on each class. In fact, the classifier corresponding to (fpr'_a, tpr'_a) has higher F-measures than the ones defined by the isometrics. This implies that the abstaining classifier exceeds the preset requirements, and therefore the classifier is clearly reliable.

Theorem 10. *If points (fpr_a, tpr_a) and (fpr_b, tpr_b) are defined by an $F_p^{c,\alpha}$-isometric and $F_n^{c,\alpha}$-isometric respectively, then the point (fpr'_a, tpr'_a) has at least the F-measures of both isometrics.*

Figure 9 gives an example where the obtained positive F-measure is approximately 5% higher than the desired performance on the positive class (i.e., than the performance of the positive isometric). The obtained negative F-measure is approximately 10% higher than the desired performance on the negative class.

***m*-estimate** An analysis of the ROC isometrics approach using m-estimate isometrics is subtle. After the transformation to an abstaining classifier we can consider that: (1) the number of a priori classified instances m is kept fixed, or (2) the parameter $\hat{m} = \frac{m}{P+N}$ is kept fixed. For brevity and readability, we will now analyze both cases in an intuitive manner. Details are found in the proofs of the theorems.

First, we consider the case where m is kept fixed. In this case upr and unr can change the distribution of a priori instances over the classes. Intuitively, if $upr < unr$ then the distribution of a priori instances in the positive m-estimate moves to the true positives resulting in a higher performance. For the negative m-estimate, the distribution moves to the false negatives resulting in a lower performance. The case of $upr > unr$ is the other way around. Therefore, an increase in performance in both classes is only possible iff $upr = unr$.

Second, we consider the case where $\hat{m}$ is kept fixed. This implies that the distribution of a priori instances over the classes is left unchanged after transformation, although absolute numbers did change. A similar reasoning as done in the first case results in an improvement of the positive m-estimate if $upr \leq unr$ and $tpr_a \geq fpr_a$. The latter condition holds for all points on the ROCCH [10]. Analogously, improvement in the negative m-estimate occurs if $upr \geq unr$ and $tpr_b \geq fpr_b$. Thus, we arrive at the following two theorems for the m-estimate.

Theorem 11. *If point (fpr_a, tpr_a) is defined by a $mest_p^{c,\hat{m}}$-isometric with $m > 0$ and if $upr \leq unr$, then the point (fpr'_a, tpr'_a) has at least the positive m-estimate of that isometric.*

Theorem 12. *If point (fpr_b, tpr_b) is defined by a $mest_n^{c,\hat{m}}$-isometric with $m > 0$ and if $upr \geq unr$, then the point (fpr'_a, tpr'_a) has at least the negative m-estimate of that isometric.*

Corollary 3. *If points (fpr_a, tpr_a) and (fpr_b, tpr_b) are defined by a $mest_p^{c,\hat{m}}$-isometric and $mest_n^{c,\hat{m}}$-isometric respectively with $m > 0$ and if $upr = unr$, then the point (fpr'_a, tpr'_a) has at least the m-estimates of both isometrics.*

From Corollary 3 it follows that a reliable classifier is obtained when the rate of abstention on the positive class and on the negative class are equal. To relax this condition, we propose to use the m-estimate for the minority class and the usual precision for the majority class. From Theorems 11 and 12 it follows that, if the minority class is the positive (negative) class, then we need an abstention characterized by $upr \leq unr$ ($upr \geq unr$). This condition is likely to be satisfied in domains with a skewed class distribution since the m-estimate isometric of the minority class has to cover a large part of ROC space in order to use sufficient data for an accurate performance indication. Figure 10 shows an example with fixed m and the negative class as minority class. The $mest_n^{c,\hat{m}}$-isometric has a low slope such that sufficient negative instances are covered for the metric to be accurate. Consequently, the condition $upr \geq unr$ is easily satisfied.

4.4 Conclusions from the Theoretical Results

The theoretical analysis of the ROC isometrics approach was divided into three cases (see Subsection 4.2). Cases 1 and 2 are trivial cases since the preset performance is at most as high as the performance of the original classifier. Case 3 is the most interesting case since it uses an abstaining classifier in order to increase the performance up to a preset value.

In the previous subsection we showed that also case 3 constructs a reliable classifier. The number of instances that it leaves unclassified is the lowest possible. The only requirement to construct the classifier is a positive isometric and a negative isometric that intersect the ROCCH.[7] We did not make any assumption on the data such as the widely-used (and often invalid) assumption of identically and independently distributed data. In addition, we did not assume that the original classifiers produce calibrated scores. Any measure of confidence can be used. Thus, from the theoretical analysis (cases 1-3) we may conclude that the ROC isometrics approach is generally applicable, and it is an effective and efficient solution to the problem statement.

5 Empirical Evaluation

In this section we provide a systematic empirical evaluation of the ROC isometrics approach (third stage). We test the approach by applying two popular classifiers on ten benchmark datasets. In Subsection 5.1 we describe the datasets and in Subsection 5.2 we explain the experimental setup. We provide our experimental results in Subsection 5.3.

[7] Note that we could also allow for situations in which only one intersection point is found. A positive (negative) isometric that does not intersect the ROCCH is represented by a threshold of plus (minus) infinity. This implies that the reliable classifier will always abstain from classifying an instance as positive or negative.

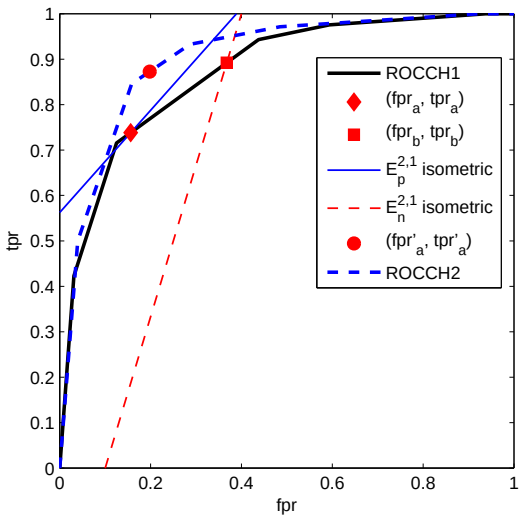


Fig. 9. Designing with F-measure isometrics: $F_p^{2,1} = 0.72$ in (fpr_a, tpr_a) and $F_n^{2,1} = 0.75$ in (fpr_b, tpr_b). The reliable classifier represented by (fpr'_a, tpr'_a) has $F_p^{1.84,1} = 0.77$ and $F_n^{1.84,1} = 0.86$. The abstention rates are $upr = 0.15$ and $unr = 0.21$.

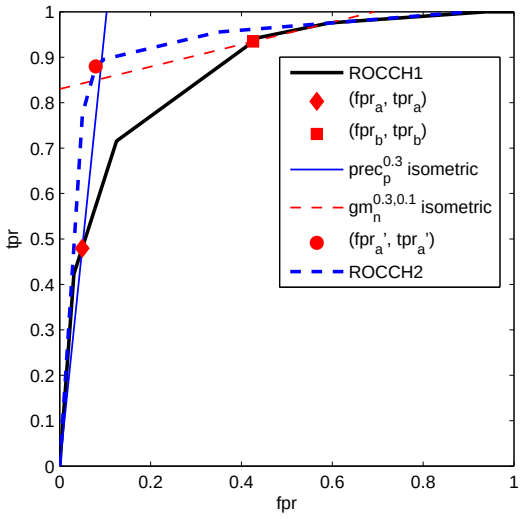


Fig. 10. Designing with precision and m-estimate isometrics: $prec_p^{0.3} = 0.97$ in (fpr_a, tpr_a) and $mest_n^{0.3,0.1} = 0.55$ in (fpr_b, tpr_b). The reliable classifier represented by (fpr'_a, tpr'_a) has $prec_p^{0.3} = 0.97$ and $mest_n^{0.34,0.18} = 0.56$. The abstention rates are $upr = 0.45$ and $unr = 0.38$.

Table 3. Benchmark datasets: name, number of instances (size), minority class, percentage of examples in the minority class, and percentage of correct classifications of k-nearest neighbour (left part) and naive Bayes (right part) as computed using 10-fold cross validation. The first six datasets are from the UCI repository, the last four are from the IJCNN competition.

name	size	min. class	% min. class	% correct classifications
heart statlog	270	p	44.44	84.81 — 85.93
house votes	342	p	34.21	94.13 — 94.49
ionosphere	350	n	35.71	90.57 — 90.29
monks3	432	n	48.15	78.70 — 83.10
sonar	208	p	46.63	89.90 — 89.82
spect	219	n	12.79	88.12 — 89.50
ada	4562	p	24.81	83.55 — 83.43
gina	3468	p	49.16	93.31 — 93.11
hiva	4229	p	3.52	97.22 — 97.89
sylva	14394	p	6.15	97.57 — 97.74

5.1 Benchmark Datasets

We tested the ROC isometrics approach on six well-known binary datasets from the UCI benchmark repository [34] and four datasets from a recent machine-learning competition [35]. The datasets vary strongly in size and in class distribution. The classes are denoted by p (positive class) and n (negative class). As a preprocessing step, all instances with missing feature values are removed as well as duplicate instances. Features are standardized to have zero mean and unit variance. Finally, for visualization purposes, linear discriminant analysis is used to project the data into a linear one-dimensional subspace of the original data space [36]. This allowed us to verify that rejected instances are instances with ambiguity in the true label (i.e., uncertain instances: instances that lie in the overlap of the score density functions of the classes). Table 3 summarizes the main characteristics of the resulting datasets.

5.2 Experimental Setup

The ROC isometrics approach is applied using ROCCHs constructed from the following two classifiers: (1) k-nearest neighbour and (2) naive Bayes. We briefly explain both classifiers and give details about the testing procedure.

Nearest Neighbour The k-nearest neighbour classifier (k-NN) classifies an instance by means of a majority vote among the labels of its k nearest neighbours ($k \geq 1$) [37]. To convert the classifier into a scoring classifier we computed the score of an instance x as follows. Let us define an ascending ordered sequence D^p with distances from instance x to its nearest neighbours with positive label. Analogously, let D^n contain ordered distances from instance x to its nearest

neighbours with negative label. The score is then defined as:

$$l(x) = \frac{\sum_{j=1}^{k} D_j^n}{\sum_{j=1}^{k} D_j^p} \ , \tag{16}$$

with subscript j representing the j-th element in a sequence [38]. Clearly, the score increases when distances to nearest neighbours of the positive class decrease and/or distances to nearest neighbours of the negative class increase.

The combination of k-NN and the ROC isometrics approach is called ROC-kNN. Nearest neighbours are found by using Euclidean distances. The number of nearest neighbours is restricted to $k = 1, 2, \ldots, 10$ and chosen in such a way that the average performance by applying ten-fold cross validation is maximized.

Naive Bayes The naive Bayes classifier (NB) is a probabilistic classifier that applies Bayes theorem with independence assumptions [39]. The score of an instance x is large when the posterior probability of the positive class is high and the posterior probability of the negative class is low. Hence, the score of instance x is defined by estimating posterior class probabilities:

$$l(x) = \frac{\hat{\mathbb{P}}(p \,|\, x)}{\hat{\mathbb{P}}(n \,|\, x)} \ , \tag{17}$$

with $\hat{\mathbb{P}}(\cdot \,|\, x)$ the conditional probability distribution over the classes for instance x as computed by the naive Bayes classifier.

The combination of NB and the ROC isometrics approach is called ROC-NB. Continuous features are handled by a simple method of binning [40]. A continuous feature is decomposed into ten discrete features of which each corresponds to an interval of values of the original feature.

Testing Procedure The classifiers ROC-kNN and ROC-NB are applied on the ten benchmark datasets in Table 3 using a ten-fold cross validation procedure. This procedure was repeated for ten times with different random permutations of the data in order to get robust results for the smaller datasets. Each training fold is used to construct an ROCCH in order to find the threshold(s). The isometrics are based on the precision measure and we set the positive precision equal to the negative precision. Hence, we measure performance in terms of accuracy (see Theorem 9). The other performance metrics considered in the article gave results that led to similar conclusions than the conclusions derived from accuracy. Therefore, in the following we generalize the description by using the term performance to denote accuracy.

Quality assessment of ROC-kNN and ROC-NB is done by measuring and reporting two key statistics of the reliable classifiers. First, we measure the average percentage of correct classifications over all test folds. This is the *empirical performance* and its value should be at least the preset performance. Second, we measure the average percentage of instances for which a label is predicted. This

is the *efficiency* and its value represents how useful the classifier is when applied in practice. A large efficiency is clearly desired.

In the experiments we consider five preset performances that we believe to be of interest in many classification tasks: 95%, 96%, 97%, 98%, and 99%. We do not consider 100% preset performance since we found that most ROCCHs have a large area under the curve although there is no vertical first line segment and/or no horizontal last line segment. This implies that at least one negative instance (positive instance) is assigned the highest (lowest) score. We therefore advice to pre-process the data to remove noise and outliers whenever the approach is used in a real-life application. For the purpose of benchmarking this is however not desired.

5.3 Experimental Results

We report and compare the experimental results of the two classifiers ROC-kNN and ROC-NB. The results are given in Tables 4 to 7. We start by giving two remarks on these tables. First, results on the `hiva` and `sylva` datasets are omitted for preset performances below 98% since this is the performance of the original classifier (see Table 3). Second, omitted values marked by an asterisk indicate that the preset performance did not result in two intersection points with the ROCCH.[8] Except for the `spect` dataset, asterisks only occur for ROC-NB which implies that the NB score is not as discriminative as the k-NN score.

Table 4 shows the empirical performances for both classifiers. We see that the empirical performances are equal to the preset performances up to statistical fluctuations, even for the small datasets. These results verify that the classification performance can be preset by the domain expert. Therefore, the two classifiers are reliable classifiers.

Table 5 shows the differences between the positive class performances and the negative class performances. A positive value indicates that the positive class performance is higher than the negative class performance. Since some datasets have a highly unbalanced class distribution, it is desired that the empirical performances on each class are approximately equal. As expected, the differences for ROC-kNN and ROC-NB show that the preset performance approximately holds for both classes, with the exception of the datasets `ionosphere`, `monks3`, and `ada`. For these datasets the classifiers seem to generate from less representative convex hull. This is especially the case for `ada`, where the sign of the difference shows that a bad performance on the positive class (the minority class) is masked by a good performance on the negative class.

Table 6 shows the efficiency (i.e., the percentage of classified instances) of ROC-kNN and ROC-NB. In general we see that the efficiency declines exponentially when the preset performance is increased. There is no clear relation between dataset characteristics and efficiency, e.g., datasets with a highly unbalanced class distribution such as `hiva` and `sylva` can still have relatively few

[8] In essence it is sufficient that at least one isometric intersects the ROCCH (see Subsection 4.4 for more details). We do not consider this setting.

unclassified instances. In addition, ROC-kNN can be claimed as the most efficient classifier since the k-NN score leads to ROCCHs that tend more toward the optimal point $(0, 1)$ in ROC space. Since different datasets and classifiers lead to different convex hulls, it is advantageous to plot the performance increase relative to the non-abstaining classifier as a function of the abstention rate. Such graphs provide a support tool for the domain expert who might like to decide on a trade-off between preset performance and abstention rate. Figure 11 shows such graphs for the `heart statlog` and `gina` datasets.

Table 7 elaborates on a practical (implementation) issue of the ROC isometrics approach. When an isometric intersects the ROCCH in an endpoint of two adjacent line segments, then the corresponding threshold can be found directly from the scores of the instances that are used to construct the ROCCH. Otherwise, a randomization of two thresholds on scores is needed, as shown in the proof of Theorem 1. For small datasets and/or large line segments this can result in a deviation from preset performance, although from Table 7 we see that in our experiments the deviations are negligible when sufficient instances are being processed.

5.4 Conclusions from the Experimental Results

Our analysis of the experimental results confirms the conclusions that we derived from the theoretical analysis. The results show that reliable classifiers can be constructed with a preset performance that is higher than the performance of the original classifier. The efficiency of such a reliable classifier depends on the shape of the original ROCCH, or in other words, the efficiency depends on the discriminative power of the scores produced by the original classifier. When a badly structured ROCCH is given as input to the approach, some preset performances may not be achievable since the isometrics do not intersect the ROCCH. In our experiments this occurred only a few times; in general, when the preset performance is almost 100%. In practice, the problem can be alleviated or avoided by pre-processing the data to detect outliers and other anomalies.

6 Comparison with Related Work

The concept of abstaining classifiers has been studied before. In this section we provide an overview of the first well-founded approach to construct abstaining classifiers and also of the approaches that are based on ROC curves. For each related approach we discuss analogies, advantages, and disadvantages with respect to the ROC isometrics approach. We will show that the approaches are extensions or generalizations of each other with the ROC isometrics approach as the most general one.

6.1 Chow's Abstaining Classifier

The first well-founded work on abstaining classifiers uses the framework of Bayesian decision theory [18]. The setting of this framework is a classification

Table 4. Empirical performances of ROC-kNN (left part) and ROC-NB (right part): empirical performances are identical to preset performances up to statistical fluctuations, even for the small datasets.

dataset	95%	96%	97%	98%	99%	95%	96%	97%	98%	99%
heart	94.5	95.6	97.0	98.0	98.8	93.5	95.0	96.1	97.5	98.0
house votes	95.8	96.1	97.2	97.7	98.7	95.0	96.4	97.0	97.7	98.7
ionosphere	95.1	95.0	96.1	97.5	98.6	94.6	95.3	*	*	*
monks3	95.6	96.0	96.9	97.7	99.0	94.6	95.3	96.7	97.1	100.0
sonar	94.8	95.9	97.0	98.1	98.8	94.1	95.9	97.3	98.1	*
spect	*	*	*	*	*	*	*	*	*	*
ada	96.0	96.9	97.8	98.5	99.3	96.0	96.9	98.4	*	*
gina	95.0	96.0	97.0	98.0	99.0	94.5	95.4	96.8	97.7	98.7
hiva	-	-	-	98.0	99.0	-	-	-	97.9	99.1
sylva	-	-	-	98.0	99.0	-	-	-	98.0	99.0

Table 5. Difference between empirical positive class and negative class performance of ROC-kNN (left part) and ROC-NB (right part): in general, empirical performances are balanced over the classes.

dataset	95%	96%	97%	98%	99%	95%	96%	97%	98%	99%
heart	0.2	1.8	-2.5	-2.5	-2.0	-1.2	-3.5	-3.7	-3.8	-6.7
house votes	-2.1	-1.7	-1.5	-0.1	-2.1	-2.2	-1.2	-1.0	-0.7	-2.8
ionosphere	-1.7	-1.9	-7.4	-10.0	-8.3	-5.7	-4.6	*	*	*
monks3	5.0	7.8	6.9	1.3	1.2	6.0	5.7	0.2	-0.5	0.0
sonar	-0.2	-0.1	-2.2	1.1	-1.3	-0.5	1.0	1.1	-0.9	*
spect	*	*	*	*	*	*	*	*	*	*
ada	-5.0	-6.5	-4.5	-5.0	-6.9	-10.1	-9.1	-7.4	*	*
gina	0.0	0.0	-0.2	-0.5	-0.6	-1.6	-1.1	-0.4	-0.3	-2.2
hiva	-	-	-	-0.6	-1.3	-	-	-	-3.4	-4.7
sylva	-	-	-	-0.2	0.1	-	-	-	-1.2	-1.1

Table 6. Efficiency of ROC-kNN (left part) and ROC-NB (right part): ROC-kNN is the most efficient reliable classifier on the majority of the datasets.

dataset	95%	96%	97%	98%	99%	95%	96%	97%	98%	99%
heart	55.6	54.0	47.0	40.0	35.0	43.3	37.0	30.0	27.4	26.3
house votes	99.4	99.4	97.6	95.9	87.9	98.5	95.9	95.6	93.3	83.0
ionosphere	84.0	78.0	52.0	35.7	30.0	80.0	74.0	*	*	*
monks3	57.2	53.3	52.8	51.2	44.2	60.5	56.7	55.4	51.9	46.3
sonar	90.0	84.5	79.5	77.5	72.5	82.5	80.5	73.5	70.0	*
spect	*	*	*	*	*	*	*	*	*	*
ada	42.6	37.8	31.9	25.2	19.6	38.1	35.3	18.5	*	*
gina	96.5	93.8	89.5	83.8	67.2	93.5	88.4	83.5	79.0	61.8
hiva	-	-	-	80.0	73.6	-	-	-	75.8	73.5
sylva	-	-	-	99.5	96.6	-	-	-	97.9	95.0

Table 7. Standard deviations of the empirical performance by applying ROC-kNN (left part) and ROC-NB (right part): randomization of thresholds has minor influence, even for the small datasets.

dataset	95%	96%	97%	98%	99%	95%	96%	97%	98%	99%
heart	0.77	0.44	1.08	0.95	1.03	1.35	2.04	1.59	1.99	1.35
house votes	0.44	0.36	0.26	0.51	0.23	0.47	0.53	0.57	0.34	0.32
ionosphere	0.33	0.33	0.42	1.35	0.40	0.74	1.12	*	*	*
monks3	0.76	0.62	0.65	0.44	0.44	0.60	0.56	0.54	0.47	0.44
sonar	0.34	0.65	0.45	0.60	0.43	0.60	0.72	0.72	0.66	*
spect	*	*	*	*	*	*	*	*	*	*
ada	0.08	0.12	0.25	0.13	0.17	0.23	0.19	0.28	*	*
gina	0.07	0.06	0.03	0.04	0.06	0.14	0.20	0.23	0.21	0.11
hiva	-	-	-	0.09	0.02	-	-	-	0.11	0.11
sylva	-	-	-	0.02	0.01	-	-	-	0.07	0.08

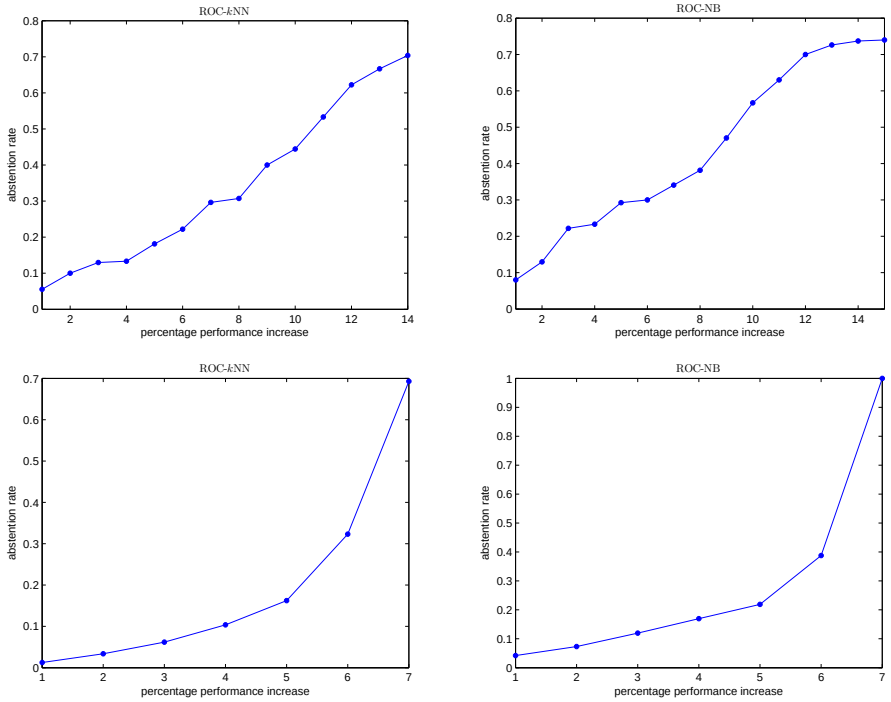


Fig. 11. Performance increase versus abstention rate curves for ROC-kNN (left part) and ROC-NB (right part): the top row shows the results for the `heart statlog` dataset and the bottom row for the `gina` dataset. Performance is increased in steps of 1% and the curves end at 100% preset performance.

problem with K classes and the expected cost as performance metric.[9] Originally, a cost matrix was required with 0 cost for correct classifications, 1 cost for incorrect classifications, and t cost for abstentions ($0 < t < 1 - 1/K$). The abstaining classifier with lowest expected cost, called Chow's abstaining classifier, predicts the class with maximum posterior probability if this value is at least $1 - t$ [19].

The classifier can be generalized as follows to include a cost matrix with arbitrary costs. Assume each instance has one of the labels $y_1, \ldots, y_K$ and represent an abstention by label $y_{K+1} = u$. We use $c(y_i, y_j)$ as the cost for predicting label y_i when the true label is y_j. The expected cost of predicting one of the labels (including the abstention option) is:

$$cost(y_i) = \sum_{j=1}^{K} c(y_i, y_j)\mathbb{P}(y_j \mid x) = \sum_{j=1}^{K} c(y_i, y_j)\frac{\pi_j \mathbb{P}(x \mid y_j)}{\mathbb{P}(x)} \quad , \tag{18}$$

with π_j the class prior, $\mathbb{P}(x \mid y_j)$ the class likelihood, and $\mathbb{P}(x)$ the marginal probability of observing instance x. The optimal decision remains the same: classify x as y_i when $cost(y_i)$ is minimal among all $cost(y_j)$ for all $j \neq i \neq u$ and $cost(y_i) < cost(u)$. The last condition implements the abstention option. For binary classification problems ($K = 2$) the abstaining classifier is given in the next theorem.

Theorem 13. *For any two-class problem with classes p and n (and extra unclassified class u), given the correct class priors and class distributions, the abstaining classifier with minimum expected cost is defined by two thresholds a and b with* $a = \frac{c(p,n)-c(u,n)}{c(u,p)} \frac{\pi_n}{\pi_p}$ *and* $b = \frac{c(n,n)}{c(n,p)-c(u,p)} \frac{\pi_n}{\pi_p}$.

Theorem 13 shows that in a two-class problem the optimal classifier according to Bayesian decision theory is an abstaining classifier with two thresholds $a > b$. A reject rule with two thresholds is a first analogy with our definition of abstaining classifiers. A second analogy with the ROC isometrics approach is the possibility for asymmetric costs. It allows to indicate that consequences of different errors or abstentions are not equally severe.

Chow's abstaining classifier has the advantage that it generalizes to multiclass problems. Abstention costs are also easily incorporated. However, its practical value is severely limited since it is assumed that the distributions of the classes are perfectly known. In practice, this is almost never the case. Careful experiments using artificially created data have shown that the classifier has expected cost significantly higher than the lowest possible expected cost when noise is added to the class distributions [41].

[9] Expected cost is equal to the error rate calculated according to a specific cost distribution. It can be seen as an alternative to our skew-sensitive accuracy metric.

6.2 Tortorella's Abstaining Classifier

Tortorella describes a second approach that chooses the two thresholds of an abstaining classifier in such a way that the expected cost is minimized [42, 22]. The approach computes the expected cost using an ROCCH.

Assume two points on the convex hull corresponding to two arbitrary thresholds $a > b$. Due to the strict ordering of these thresholds we have that a true positive for the classifier corresponding to a is a true positive for the classifier corresponding to b. Analogously, a true negative for the classifier corresponding to b is a true negative for the classifier corresponding to a. So indeed, when the classifiers disagree, we actually abstain from classification. The expected cost can therefore be defined in terms of the four performance characteristics:

$$\begin{aligned} cost(a,b) = \pi_p c(n,p) fnr_b + \pi_n c(n,n) tnr_b + \\ \pi_p c(p,p) tpr_a + \pi_n c(p,n) fpr_a + \\ \pi_p c(u,p) upr + \pi_n c(u,n) unr \ . \end{aligned} \tag{19}$$

Rewriting Eq. 19 results in a new equation in which the optimal values for a and b can be separately evaluated. As computed in [22] each of these evaluations defines two sets of lines in ROC space with slopes, namely:

$$s_a = \frac{c(p,n) - c(u,n)}{c(u,p)} \frac{\pi_n}{\pi_p} \ , \text{ and} \tag{20}$$

$$s_b = \frac{c(u,n)}{c(n,p) - c(u,p)} \frac{\pi_n}{\pi_p} \ . \tag{21}$$

Among the lines with slope s_a we can find the optimal a as the threshold value corresponding to the point in which these lines touch the ROCCH (or coincide with exactly one line segment). Analogously, we can find the optimal b using the lines with slope s_b.

Tortorella's abstaining classifier is equal to that of Chow when there is complete knowledge of the distributions of the classes [24]. In addition, Tortorella's approach can be applied when complete knowledge is not available since the optimal thresholds depend solely on the cost matrix and the ROCCH. Consequently, the approach leads to abstaining classifiers that have in general lower expected cost than that of Chow when trained under the same conditions.

Compared with the ROC isometrics approach, there is a clear analogy since both approaches rely on the slope of line segments of the ROCCH. A disadvantage of Tortorella's abstaining classifier is that it can only deal with the expected cost as performance metric. A second disadvantage is that it requires knowledge of the costs for abstention. The optimal thresholds of the classifier have been shown to be extremely sensitive to small changes in the abstention costs. In an experimental study on 15 datasets it was shown that an increase from abstention cost 0.1 to 0.2 resulted in an average decrease of 26% in unclassified instances and a decrease of 21% in cost improvement [23]. So, a small change in abstention

costs can lead to a drastic change in abstention rate and performance improvement. It is also not easy to determine the relation between abstention costs and misclassification costs. Thus, an approach that explicitly needs a full cost matrix has limited practical purposes. The ROC isometrics approach is more natural as discussed in the next subsection.

6.3 Pietraszek's Abstaining Classifier

The approach in the previous subsection has been extended by Pietraszek in order to handle unknown abstention costs [23]. He introduced two enhanced approaches in which either the expected cost or the abstention rate is bounded by the domain expert.

Pietraszek's first abstaining classifier has the same motivation as we used for the ROC isometrics approach: guarantee a minimal preset performance while allowing as few abstentions as possible. A constraint on the minimal performance is easily quantifiable and intuitive for the domain expert. Hence, reliable classifiers can be constructed that meet predefined (sometimes contractual) norms. A first disadvantage of Pietraszek's abstaining classifier when compared to the ROC isometrics approach is that it has no closed-form solution in the general case. Therefore, a specifically designed algorithm was presented in case of the expected cost. The generalization to other performance metrics is not trivial. A second disadvantage is the lack of a clear geometrical interpretation and visualization of the influence of different preset performances and skew ratios.

Pietraszek's second abstaining classifier is the inverse of the first one: find the classifier with minimum expected cost among all classifiers with abstention rate as most as high as a preset value. This approach is natural in resource-constrained situations where a human expert can only handle a limited number of events in some time period. The abstaining classifier decides automatically which instances should be passed to the human for verification. Its usefulness has been argued and experimentally validated for the application of intrusion detection [43]. Unfortunately, the ROC isometrics approach cannot be directly used for these situations since the abstention rate is known only after construction of the isometrics. Yet, by plotting a graph depicting preset performance versus abstention rate, much insight is obtained. In this way the domain expert has a support tool to help decide what is the best trade-off between performance improvement and the percentage of unclassified instances. Note that plotting such a graph is computationally inexpensive. In addition, the ROC isometrics approach can make such a graph for different performance metrics. The restriction to expected cost in Pietraszek's second classifier is a disadvantage for some domains.

7 Conclusions

As stated in the introduction, many domains face the problem of deciding to what extent classifiers can be applied when incorrect classifications may have

severe consequences. Therefore, we formulated the following problem statement: can we develop a feasible approach by which a classifier is constructed that guarantees a preset classification performance on each class? If the answer is affirmative then such a classifier is a reliable classifier, ready for safe employment in the domain.

In the article we presented the ROC isometrics approach as an effective solution to the problem statement. The approach is a successful combination of: (1) classifier visualization in ROC space, (2) isometrics, and (3) an abstaining classifier that uses two thresholds to identify uncertain instances. We remark that abstaining classifiers are a general and intuitive approach to improve the classification performance since in many domains abstaining classifiers are preferred over classifiers who always make a prediction but are too often wrong. We provided an analysis of the approach when precision, F-measure, and m-estimate are used to measure classification performance. In addition, our analysis showed that the approach can also be used with a cost-sensitive version of accuracy as well as with any combination of the aforementioned performance metrics. We empirically tested the approach by applying two popular classifiers on ten benchmark datasets. The experimental results confirmed the formal analysis. Comparison with related work clearly revealed the contributions of the new approach.

Based on our theoretical analysis, the empirical evaluation, and our discussion about related work, we are able to formulate four conclusions. First, we may conclude that the ROC isometrics approach is generally applicable since any classifier can be used to construct an ROC curve. Some classifiers such as naive Bayes and neural networks naturally provide scores. For other classifiers, such as nearest neighbour and support vector machine, a postprocessing technique may be needed or preferred [44–46]. Second, we may conclude that the ROC isometrics approach does not commit to specific cost distributions and class distributions since skew-sensitive isometrics are used. When there is a change in these distributions, a new reliable classifier is constructed efficiently from the original ROCCH by two straightforward steps: (1) an update of the skew ratio in the isometrics, and (2) a recalculation of the intersection points and thresholds. Third, we may conclude that the ROC isometrics approach is clearly efficient in terms of time complexity since it only involves the computation of intersection points. It is also efficient in terms of space complexity since it solely involves storing a convex hull and a short list of thresholds. Fourth, we may conclude that the approach is easy-to-visualize. In summary, we state that the ROC isometrics approach provides a satisfactory answer to the problem statement and opens new grounds for immediate applications in domains with high error costs such as the domain of law enforcement and medical diagnosis.

Acknowledgments

We thank the reviewers for useful comments and suggestions. The first author is supported by the Dutch Organization for Scientific Research (NWO), ToKeN programme, viz. the IPOL project, grant nr: 634.000.435.

References

1. Vanderlooy, S., Sprinkhuizen-Kuyper, I., Smirnov, E.: Reliable classifiers in ROC space. In Saeys, Y., Tsiporkova, E., Baets, B.D., de Peer, Y.V., eds.: Proceedings of the 15th Annual Machine Learning Conference of Belgium and the Netherlands (BENELEARN 2006), Ghent, Belgium (2006) 113–120
2. Vanderlooy, S., Sprinkhuizen-Kuyper, I., Smirnov, E.: An analysis of reliable classifiers through ROC isometrics. In Lachiche, N., Ferri, C., Macskassy, S., eds.: Proceedings of the ICML 2006 Workshop on ROC Analysis (ROCML 2006), Pittsburgh, USA (2006) 55–62
3. Lachiche, N., Flach, P.: Improving accuracy and cost of two-class and multi-class probabilistic classifiers using ROC curves. In Fawcett, T., Mishra, N., eds.: Proceedings of the 20th International Conference on Machine Learning (ICML 2003), Washington, DC, USA, AAAI Press (2003) 416–423
4. Swets, J., Dawes, R., Monahan, J.: Better decisions through science. Scientific American **283**(4) (2000) 82–87
5. Fawcett, T.: ROC graphs: Notes and practical considerations for researchers. Technical Report HPL-2003-4, HP Laboratories (2003)
6. Provost, F., Fawcett, T.: Robust classification for imprecise environments. Machine Learning **42**(3) (2001) 203–231
7. Provost, F., Fawcett, T.: Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. In Heckerman, D., Mannila, H., Pregibon, D., eds.: Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining (KDD 1997), AAAI Press (1997) 43–48
8. Ling, C., Huang, J., Zhang, H.: AUC: a statistically consistent and more discriminating measure than accuracy. In Gottlob, G., Walsh, T., eds.: Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI 2003), AAAI Press (2003)
9. Provost, F., Fawcett, T., Kohavi, R.: The case against accuracy estimation for comparing induction algorithms. In Shavlik, J., ed.: Proceedings of the 15th International Conference on Machine Learning (ICML 1998), Madison, Wisconsin, USA, Morgan Kaufmann (1998) 43–48
10. Flach, P., Wu, S.: Repairing concavities in ROC curves. In Kaelbling, L., Saffiotti, A., eds.: Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI 2005), Edinburgh, UK, Professional Book Center (2005) 702–707
11. Fawcett, T., Flach, P.: A response to Webb and Ting's on the application of ROC analysis to predict classification performance under varying class distributions. Machine Learning **58**(1) (2005) 33–38
12. Ting, K.: A study of the effect of class distribution using cost-sensitive learning. In Lange, S., Satoh, K., Smith, C., eds.: Proceedings of the 5th International Conference on Discovery Science (DS 2002), Lübeck, Germany, Springer (2002) 98–112
13. Weiss, G., Provost, F.: Learning when training data are costly: The effect of class distribution on tree induction. Journal of Artificial Intelligence Research **19** (2003) 315–354
14. Weiss, G., McCarthy, K., Zabar, B.: Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error costs? In Stahlbock, R., Crone, S., Lessmann, S., eds.: Proceedings of the 2007 International Conference on Data Mining (DMIN 2007), Las Vegas, NV, USA, CSREA Press (2007) 35–41

15. Elkan, C.: The foundations of cost-sensitive learning. In Nebel, B., ed.: Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI 2001), Seattle, WA, USA, Morgan Kaufmann (2001) 973–978
16. Flach, P.: The geometry of ROC space: Understanding machine learning metrics through ROC isometrics. In Fawcett, T., Mishra, N., eds.: Proceedings of the 20th International Conference on Machine Learning (ICML 2003), Washington, DC, USA, AAAI Press (2003) 194–201
17. Friedel, C., Rückert, U., Kramer, S.: Cost curves for abstaining classifiers. In Lachiche, N., Ferri, C., Macskassy, S., eds.: Proceedings of the ICML 2006 Workshop on ROC Analysis (ROCML 2006), Pittsburgh, PA, USA (2006) 33–40
18. Chow, C.: An optimum character recognition system using decision functions. IRE Transactions on Electronic Computers **6**(4) (1957) 247–254
19. Chow, C.: On optimum recognition error and reject tradeoff. IEEE Transactions on Information Theory **16**(1) (1970) 41–46
20. Pazzani, M., Murphy, P., Ali, K., Schulenburg, D.: Trading off coverage for accuracy in forecasts: Applications to clinical data analysis. In: Working Notes of the AAAI Symposium on AI in Medicine, Stanford, CA, USA, AAAI Press (1994) 106–110
21. Ferri, C., Hernández-Orallo, J.: Cautious classifiers. In Hernández-Orallo, J., Ferri, C., Lachiche, N., Flach, P., eds.: Proceedings of the 1st International Workshop on ROC Analysis in Artificial Intelligence (ROCAI 2004), Valencia, Spain (2004) 27–36
22. Tortorella, F.: A ROC-based reject rule for dichotomizers. Pattern Recognition Letters **26**(2) (2005) 167–180
23. Pietraszek, T.: On the use of ROC analysis for the optimization of abstaining classifiers. Machine Learning **68**(2) (2007) 137–169
24. Santos-Pereira, C., Pires, A.: On optimal reject rules and ROC curves. Pattern Recognition Letters **26**(7) (2005) 943–952
25. Muzzolini, R., Yang, Y.H., Pierson, R.: Classifier design with incomplete knowledge. Pattern Recognition **31**(4) (1998) 345–369
26. Frélicot, C., Mascarilla, L.: Reject strategies driven combination of pattern classifiers. Pattern Analysis and Applications **5**(2) (2002) 234–243
27. Ferri, C., Flach, P., Hernández-Orallo, J.: Delegating classifiers. In Hernández-Orallo, J., Ferri, C., Lachiche, N., Flach, P., eds.: Proceedings of the 1st International Workshop on ROC Analysis in Artificial Intelligence (ROCAI 2004), Valencia, Spain (2004) 37–44
28. Fürnkranz, J., Flach, P.: Roc 'n' rule learning – towards a better understanding of covering algorithms. Machine Learning **58**(1) (2005) 39–77
29. Rijsbergen, C.V.: Information Retrieval. 2nd edn. Department of Computer Science, University of Glasgow (1979)
30. Džeroski, S., Bratko, I.: Handling noise in inductive logic programming. In Muggleton, S., Furukawa, K., eds.: Proceedings of the 2nd International Workshop on Inductive Logic Programming (ILP 1992), Tokyo, Japan, Institute for New Generation Computer Technology (1992) 109–125
31. Lavrac, N., Dzeroski, S.: Inductive Logic Programming: Techniques and Applications. Ellis Horwood, New York, NY, USA (1994)
32. Macskassy, S., Provost, F.: Confidence bands for ROC curves: Methods and an empirical study. In Hernández-Orallo, J., Ferri, C., Lachiche, N., Flach, P., eds.: Proceedings of the 1st International Workshop on ROC Analysis in Artificial Intelligence (ROCAI 2004), Valencia, Spain (2004) 61–70

33. Agarwal, S., Graepel, T., Herbrich, R., Har-Peled, S., Roth, D.: Generalization error bounds for the area under the ROC curve. Journal of Machine Learning Research **6**(Apr) (2005) 393–425
34. Newman, D., Hettich, S., Blake, C., Merz, C.: UCI repository of machine learning databases (1998)
35. Guyon, I.: Data representation discovery workshop of the 20th international joint conference on neural networks (IJCNN 2007) (2007)
36. Ye, J., Janardan, R., Li, Q.: Two-dimensional linear discriminant analysis. In: Advances in Neural Information Processing Systems 17 (NIPS 2004), Vancouver, British Columbia, Canada (2004) 1569–1576
37. Cover, T., Hart, P.: Nearest neighbor pattern classification. IEEE Transactions on Information Theory **13**(1) (1967) 21–27
38. Proedrou, K., Nouretdinov, I., Vovk, V., Gammerman, A.: Transductive confidence machines for pattern recognition. Technical Report 01-02, Royal Holloway University of London, London, UK (2001)
39. Domingos, P., Pazzani, M.: On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning **29**(2) (1997) 103–130
40. Lu, J., Yang, Y., Webb, G.: Incremental discretization for naive Bayes classifier. In Li, X., Zaïane, O., Li, Z., eds.: Proceedings of the 2nd International Conference on Advanced Data Mining and Applications (ADMA 2006), Xi'an, China, Springer (2006) 223–238
41. Marrocco, C., Molinara, M., Tortorella, F.: An empirical comparison of ideal and empirical ROC-based reject rules. In Perner, P., ed.: Proceedings of the 5th International Conference on Machine Learning and Data Mining in Pattern Recognition (MLDM 2007), Leipzig, Germany, Springer (2007) 47–60
42. Tortorella, F.: Reducing the classification cost of support vector classifiers through an ROC-based reject rule. Pattern Analysis and Applications **7**(2) (2004) 128–143
43. Pietraszek, T.: Classification of intrusion detection alerts using abstaining classifiers. Intelligent Data Analysis **11**(3) (2007) 293–316
44. Cestnik, B.: Estimating probabilities: A crucial task in machine learning. In Aiello, L., ed.: Proceedings of the 9th European Conference on Artificial Intelligence (ECAI 1990), Stockholm, Sweden, Pitman Publishing (1990) 147–149
45. Atiya, A.: Estimating the posterior probabilities using the k-nearest neighbor rule. Neural Computation **17**(3) (2005) 731–740
46. Niculescu-Mizil, A., Caruana, R.: Predicting good probabilities with supervised learning. In: Proceedings of the 22nd International Conference on Machine Learning (ICML 2005), Bonn, Germany, ACM Press (2005) 625–632

A Proofs

This appendix is subdivided into 13 smaller appendices. They contain the proofs of the theorems. The number of smaller appendices and the theorems coincide. We start by providing some information for the proofs.

For the proof of the convexity of the abstention ROC curve (Theorem 3), we note that an ROC curve is convex if for all points (fpr_1, tpr_1), (fpr_2, tpr_2), and (fpr_3, tpr_3) such that $0 \leq fpr_1 < fpr_2 < fpr_3 \leq 1$, it holds that:

$$tpr_2 \geq \frac{fpr_3 - fpr_2}{fpr_3 - fpr_1} tpr_1 + \frac{fpr_2 - fpr_1}{fpr_3 - fpr_1} tpr_3 \ , \tag{22}$$

or equivalently:

$$\frac{tpr_2 - tpr_1}{fpr_2 - fpr_1} \geq \frac{tpr_3 - tpr_2}{fpr_3 - fpr_2} \ . \tag{23}$$

For the proofs of the dominance relations between the original ROCCH and the abstention ROCCH (Theorems 4 - 6), we denote the points on the original ROCCH by $(x, tpr(x))$. So, we introduce the function $tpr(x)$ for the highest true positive rate corresponding to a false positive rate x. Analogously, the points $(x, tpr'(x))$ are the points on the abstention ROCCH.

For the proofs of the effectiveness of the ROC isometrics approach (Theorems 8 - 12), we assume for simplicity that $c = \frac{N}{P}$. After transformation to an abstaining classifier, a new skew ratio is obtained that we denote by $c' = c\frac{1-unr}{1-upr}$.

A.1 Proof of Theorem 1

We fix an arbitrary point (fpr_i, tpr_i) on an ROCCH and denote the corresponding classifier as c_i. The line segment of the convex hull that contains this point is denoted by L_i. We can distinguish two cases.

- **Case 1:** (fpr_i, tpr_i) is an endpoint of L_i.
  Endpoints of line segments of an ROCCH are points that also lie on the ROC curve itself. Hence, thresholds for the classifiers corresponding to these points are found directly from the scores of the instances that are used to construct the ROC curve.
- **Case 2:** (fpr_i, tpr_i) is not an endpoint of L_i.
  Denote the classifiers corresponding to the endpoints of L_i by c_1 and c_2. We define $d_{1,i} = fpr_i - fpr_1$ and $d_{1,2} = fpr_2 - fpr_1$. For each instance x the classification of c_i is a random variable that takes either of values of $c_1(x)$ or $c_2(x)$ with probabilities:

$$\mathbb{P}(c_i(x) = c_1(x)) = 1 - \frac{d_{1,i}}{d_{1,2}} \text{ , and}$$
$$\mathbb{P}(c_i(x) = c_2(x)) = \frac{d_{1,i}}{d_{1,2}} \ .$$

  It follows that the expectation of fpr is equal to fpr_i:

$$\begin{aligned} E[fpr] &= \left(1 - \frac{d_{1,x}}{d_{1,2}}\right) fpr_1 + \frac{d_{1,x}}{d_{1,2}} fpr_2 \\ &= \frac{(fpr_2 - fpr_1)d_{1,x}}{d_{1,2}} + fpr_1 \\ &= fpr_i \ . \end{aligned}$$

  Analogously, we can show that the expectation of tpr is equal to tpr_i. So, again we can construct classifier c_i.

□

A.2 Proof of Theorem 2

The classifier represented by a point (fpr_i, tpr_i) between the points $(0,0)$ and (fpr_a, tpr_a) on the original ROC curve classifies $tpr_i P$ instances correctly as positive and $fpr_i N$ instances incorrectly as positive. The classifier of the point (fpr'_i, tpr'_i) on the abstention ROC curve classifies $tpr'_i(P - UP)$ instances correctly as positive and $fpr'_i(N - UN)$ instances incorrectly as positive. Since these instances are the same we have:

$$\begin{aligned} fpr_i N &= fpr'_i(N - UN) \text{ , and} \\ tpr_i P &= tpr'_i(P - UP) \text{ ,} \end{aligned}$$

and straightforward rewriting results in Eq. 9.

Analogously, the classifier represented by a point (fpr_i, tpr_i) between the points (fpr_b, tpr_b) and $(1,1)$ on the original ROC curve classifies $(1 - fpr_i)N$ instances correctly as negative and $(1 - tpr_i)P$ instances incorrectly as negative. The classifier of the point (fpr'_i, tpr'_i) on the abstention ROC curve classifies $(1 - fpr'_i)(N - UN)$ instances correctly as negative and $(1 - tpr'_i)(P - UP)$ instances incorrectly as negative. Since these instances are the same we have:

$$\begin{aligned} (1 - fpr_i)N &= (1 - fpr'_i)(N - UN) \text{ , and} \\ (1 - tpr_i)P &= (1 - tpr'_i)(P - UP) \text{ ,} \end{aligned}$$

and rewriting results in Eq. 10. □

A.3 Proof of Theorem 3

For the abstention ROC curve to be convex, we need to show that either Eq. 22 or Eq. 23 holds for all points (fpr'_1, tpr'_1), (fpr'_2, tpr'_2), and (fpr'_3, tpr'_3) such that $0 \leq fpr'_1 < fpr'_2 < fpr'_3 \leq 1$. We split our proof into five different cases.

- **Case 1:** $0 \leq fpr'_1 < fpr'_2 < fpr'_3 \leq fpr_a$.
  The use of Eq. 9 shows that Eq. 23 for the abstention ROC curve is equivalent to:
  $$\frac{(tpr_2 - tpr_1)(1 - unr)}{(fpr_2 - fpr_1)(1 - upr)} \geq \frac{(tpr_3 - tpr_2)(1 - unr)}{(fpr_3 - fpr_2)(1 - upr)} \text{ ,}$$
  and this holds because of the convexity of the original ROC curve.
- **Case 2:** $fpr'_b \leq fpr'_1 < fpr'_2 < fpr'_3 \leq 1$.
  The proof of this case is similar to that of case 1.
- **Case 3:** $0 \leq fpr'_1 < fpr'_2 = fpr'_a = fpr'_b < fpr'_3 \leq 1$.
  The analogue of Eq. 23 for the abstention ROC curve is:
  $$\frac{tpr'_a - tpr'_1}{fpr'_a - fpr'_1} \geq \frac{tpr'_3 - tpr'_b}{fpr'_3 - fpr'_b} \text{ ,} \tag{24}$$
  and can be rewritten as:
  $$\frac{(tpr_a - tpr_1)(1 - unr)}{(fpr_a - fpr_1)(1 - upr)} \geq \frac{(tpr_3 - tpr_b)(1 - unr)}{(fpr_3 - fpr_b)(1 - upr)} \text{ ,} \tag{25}$$

using Eq. 9 and Eq. 10 for the left hand side and the right hand side of Eq. 24, respectively. Equation 25 holds because of the convexity of the original ROC curve:

$$\frac{tpr_a - tpr_1}{fpr_a - fpr_1} \geq \frac{tpr_b - tpr_a}{fpr_b - fpr_a} \geq \frac{tpr_3 - tpr_b}{fpr_3 - fpr_b} \ .$$

– **Case 4:** $0 \leq fpr'_1 < fpr'_2 < fpr'_a = fpr'_b < fpr'_3 \leq 1$.
We write analogues of Eq. 22 for the abstention ROC curve. We know from case 1 that:

$$tpr'_2 \geq \frac{fpr'_a - fpr'_2}{fpr'_a - fpr'_1} tpr'_1 + \frac{fpr'_2 - fpr'_1}{fpr'_a - fpr'_1} tpr'_a \ , \tag{26}$$

and from case 3 that:

$$tpr'_a \geq \frac{fpr'_3 - fpr'_a}{fpr'_3 - fpr'_2} tpr'_2 + \frac{fpr'_a - fpr'_2}{fpr'_3 - fpr'_2} tpr'_3 \ . \tag{27}$$

The use of Eq. 27 to substitute tpr'_a in Eq. 26 results in:

$$tpr'_2 \geq \frac{fpr'_3 - fpr'_2}{fpr'_3 - fpr'_1} tpr'_1 + \frac{fpr'_2 - fpr'_1}{fpr'_3 - fpr'_1} tpr'_3 \ .$$

– **Case 5:** $0 \leq fpr'_1 < fpr'_a = fpr'_b < fpr'_2 < fpr'_3 \leq 1$.
The proof of this case is similar to that of case 4. □

A.4 Proof of Theorem 4

The abstention ROCCH dominates the original one between the points $(0, 0)$ and (fpr_a, tpr_a) when $tpr'(x) \geq tpr(x)$ for $0 \leq x \leq fpr_a$. The part of the original ROCCH from the point $(0, 0)$ to the point (fpr_a, tpr_a) is transformed into the abstention ROCCH according to:

$$(x, tpr(x)) \rightarrow \left(\frac{x}{1 - unr}, \frac{tpr(x)}{1 - upr} \right) \ .$$

It follows that:

$$tpr' \left(\frac{x}{1 - unr} \right) = \frac{tpr(x)}{1 - upr} \ ,$$

or equivalently:

$$tpr'(x) = \frac{tpr\left(x(1 - unr)\right)}{1 - upr} \ .$$

Now we see that $tpr'(x) \geq tpr(x)$ for $0 \leq x \leq fpr_a$ iff:

$$tpr(x(1 - unr)) \geq (1 - upr) tpr(x) \ . \tag{28}$$

To proof Eq. 28 we use the fact that the original ROC curve is convex between the points $(0,0)$ and $(x, tpr(x))$. Hence, Eq. 22 for $fpr_1 = 0$, $fpr_2 = x(1-unr)$, and $fpr_3 = x$ reads:

$$tpr(x(1-unr)) \geq \frac{x - x(1-unr)}{x} 0 + \frac{x(1-unr)}{x} tpr(x) \ . \tag{29}$$

From a direct comparison between Eqs. 28 and 29 we see that Eq. 28 is true if $unr \leq upr$. □

A.5 Proof of Theorem 5

The proof is similar to that of Theorem 4. □

A.6 Proof of Theorem 6

Note that $tpr(0) = tpr_0 > 0$. We start the proof with Eq. 28 that we derived in the proof of Theorem 4. Since the original ROC curve is convex between the points $(0, tpr_0)$ and $(x, tpr(x))$ for $0 \leq x \leq fpr_a$, Eq. 22 for $fpr_1 = 0$, $fpr_2 = x(1-unr)$, and $fpr_3 = x$ results in:

$$tpr(x(1-unr)) \geq unr \ tpr_0 + (1-unr) tpr(x) \ .$$

Thus, Eq. 28 holds iff:

$$unr \ tpr_0 + (1-unr) tpr(x) \geq (1-upr) tpr(x) \ ,$$

or equivalently iff:

$$tpr_0 \geq \left(1 - \frac{upr}{unr}\right) tpr(x) \ .$$

This holds when $unr \leq upr$ (cf. Theorem 4), and when both $unr > upr$ and $tpr(x) \leq \frac{1}{1-\frac{upr}{unr}} tpr_0$. The last inequality holds for the whole interval $0 \leq x \leq fpr_a$ iff $tpr_a \leq \frac{1}{1-\frac{upr}{unr}} tpr_0$.

From Theorem 5 and the result above it follows that the abstention ROCCH dominates the original ROCCH on the covered part if $tpr_0 > 0$, $unr > upr$, and if $tpr_a \leq \frac{1}{1-\frac{upr}{unr}} tpr_0$. □

A.7 Proof of Theorem 7

The proof is similar to that of Theorem 6. □

A.8 Proof of Theorem 8

The positive precisions in (fpr_a, tpr_a) and (fpr'_a, tpr'_a) are defined as follows:

$$prec_p^c (fpr_a, tpr_a) = \frac{tpr_a}{tpr_a + c \, fpr_a} \ , \text{ and} \tag{30}$$

$$prec_p^{c'} \left(fpr'_a, tpr'_a\right) = \frac{tpr'_a}{tpr'_a + c' \, fpr'_a} \ . \tag{31}$$

Substitution of Eq. 9 in Eq. 31 results in Eq. 30. Analogously, we can use Eq. 10 to show that the negative precisions in (fpr_b, tpr_b) and (fpr'_b, tpr'_b) are equal. The theorem follows since $(fpr'_b, tpr'_b) = (fpr'_a, tpr'_a)$. □

A.9 Proof of Theorem 9

Since the positive precision and the negative precision in (fpr'_a, tpr'_a) are equal, we can define $a = prec_p^{c'} = prec_n^{c'}$ and rewrite the equations of the two variants of the precision metric to obtain:

$$tpr'_a = a\left(tpr'_a + c'\ fpr'_a\right) \text{ , and}$$
$$tnr'_a = a\left(tnr'_a + \frac{1}{c'} fnr'_a\right) \ .$$

It follows that:

$$tpr'_a + c'\ tnr'_a = a\left(tpr'_a + c'\ fpr'_a + c'\ tnr'_a + fnr'_a\right) \ ,$$

or equivalently:

$$a = \frac{tpr'_a + c'\ tnr'_a}{tpr'_a + c'\ fpr'_a + c'\ tnr'_a + fnr'_a} \ . \tag{32}$$

Equation 32 is the skew-sensitive version of accuracy with skew ratio c'. □

A.10 Proof of Theorem 10

The positive F-measures in (fpr_a, tpr_a) and (fpr'_a, tpr'_a) are defined as follows:

$$F_p^{c,\alpha}(fpr_a, tpr_a) = \frac{\left(1+\alpha^2\right) tpr_a}{\alpha^2 + tpr_a + c\ fpr_a} \text{ , and} \tag{33}$$
$$F_p^{c',\alpha}\left(fpr'_a, tpr'_a\right) = \frac{\left(1+\alpha^2\right) tpr'_a}{\alpha^2 + tpr'_a + c'\ fpr'_a} \ . \tag{34}$$

Using Eq. 9 we can rewrite Eq. 34 to obtain:

$$F_p^{c',\alpha}\left(fpr'_a, tpr'_a\right) = \frac{\left(1+\alpha^2\right) tpr_a}{\alpha^2(1-upr) + tpr_a + c\ fpr_a} \ .$$

It follows that $F_p^{c',\alpha}\left(fpr'_a, tpr'_a\right) > F_p^{c,\alpha}(fpr_a, tpr_a)$ since $0 < upr < 1$. Analogously, we can use Eq. 10 to show that the negative F-measure in (fpr'_b, tpr'_b) is higher than the negative F-measure in (fpr_b, tpr_b). The theorem follows since $(fpr'_b, tpr'_b) = (fpr'_a, tpr'_a)$. □

A.11 Proof of Theorem 11

The positive m-estimates in (fpr_a, tpr_a) and (fpr'_a, tpr'_a) are defined as follows:

$$mest_p^{c,\hat{m}}(fpr_a, tpr_a) = \frac{tpr + \hat{m}}{tpr + c\, fpr + \hat{m}(1+c)} \text{ , and} \tag{35}$$

$$mest_p^{c',\hat{m}'}\left(fpr'_a, tpr'_a\right) = \frac{tpr' + \hat{m}'}{tpr' + c'\, fpr' + \hat{m}'(1+c')} \ . \tag{36}$$

As explained in the main text, two cases are distinguished dependent on what we leave unchanged after transformation to an abstaining classifier. We now proof the two possible cases.

- **Case 1:** m is not changed after transformation.
  In this case we have:

  $$\hat{m}' = \frac{m}{P(1-upr) + N(1-unr)} \ . \tag{37}$$

  Substitution of Eqs. 9 and 37 in Eq. 36 followed by a straightforward calculation results in:

  $$mest_p^{c',\hat{m}'}\left(fpr'_a, tpr'_a\right) = \frac{tpr + m\frac{1-upr}{P(1-upr)+N(1-unr)}}{tpr + c\, fpr + \hat{m}(1+c)} \ . \tag{38}$$

  From a direct comparison between Eq. 35 and Eq. 38 it follows that the claim $mest_p^{c',\hat{m}'}\left(fpr'_a, tpr'_a\right) \geq mest_p^{c,\hat{m}}(fpr_a, tpr_a)$ holds iff:

  $$\frac{1-upr}{P(1-upr) + N(1-unr)} \geq \frac{1}{P+N} \ .$$

  This is true iff $upr \leq unr$.
- **Case 2:** $\hat{m}$ is not changed after transformation.
  In this case we have $\hat{m}' = \hat{m}$. Substitution of Eq. 9 in Eq. 36 followed by a straightforward calculation results in:

  $$mest_p^{c',\hat{m}'}\left(fpr'_a, tpr'_a\right) = \frac{tpr + \hat{m}(1-upr)}{tpr + c\, fpr + \hat{m}(1-upr+c(1-unr))} \ . \tag{39}$$

  From a direct comparison between Eq. 35 and Eq. 39 it follows that the claim $mest_p^{c',\hat{m}}\left(fpr'_a, tpr'_a\right) \geq mest_p^{c,\hat{m}}(fpr_a, tpr_a)$ holds iff:

  $$\hat{m}(unr - upr) + (tpr_a\ unr - fpr_a\ upr) \geq 0 \ .$$

  This is true if $upr \leq unr$ and $tpr_a \geq fpr_a$.

□

A.12 Proof of Theorem 12

The proof is similar to that of Theorem 11. □

A.13 Proof of Theorem 13

Trivially, let the cost of abstention be in between the costs of correct classification and incorrect classification. The score of an instance that is used to build an ROC curve is an estimate of the likelihood ratio $L = \mathbb{P}(x \mid p)/\mathbb{P}(x \mid n)$.

Suppose the best decision is to abstain from classifying an instance x. This means that the following two conditions are satisfied: $cost(p) > cost(u)$ and $cost(n) > cost(u)$. A straightforward derivation of the first condition gives:

$$\begin{aligned} c(p,n)\mathbb{P}(n \mid x) &> c(u,p)\mathbb{P}(p \mid x) + c(u,n)\mathbb{P}(n \mid x) \\ c(p,n)\pi_n &> c(u,p)\pi_p L + c(u,n)\pi_n \\ L &< \frac{c(p,n) - c(u,n)}{c(u,p)} \frac{\pi_n}{\pi_p} = a \end{aligned} \tag{40}$$

Similarly, we can derive that the second condition is equal to:

$$L > \frac{c(u,n)}{c(n,p) - c(u,p)} \frac{\pi_n}{\pi_p} = b \tag{41}$$

Combining Eqs. 40 and 41 shows that we abstain from classifying instance x when $b < L < a$. If this is not the case then we use the standard decision rule: classify as positive when $L \geq a$ and as negative as $L \leq b$. □